



(51) International Patent Classification:

G06F 16/35 (2019.01) G06K 9/62 (2006.01)
G06F 40/279 (2020.01) G06N 3/08 (2006.01)

(21) International Application Number:

PCT/US2022/024995

(22) International Filing Date:

15 April 2022 (15.04.2022)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/177,406 21 April 2021 (21.04.2021) US
17/720,431 14 April 2022 (14.04.2022) US

(71) Applicant: **NEC LABORATORIES AMERICA, INC.**
[US/US]; 4 Independence Way, Suite 200, Princeton, New
Jersey 08540 (US).

(72) Inventors: **LIU, Yanchi**; 34 Beechwood Ct, -, Monmouth
Junction, New Jersey 08852 (US). **CHEN, Haifeng**; 11
Blackhawk Ct, -, West Windsor, New Jersey 08550 (US).
ZHANG, Xuchao; 7300 Summit Rock Rd, -, Elkridge,
Maryland Elkridge (US).

(74) Agent: **BITETTO, James J.**; 401 Broadhollow Rd, -,
Melville, New York 11747 (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, 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: TECHNICAL SPECIFICATION MATCHING

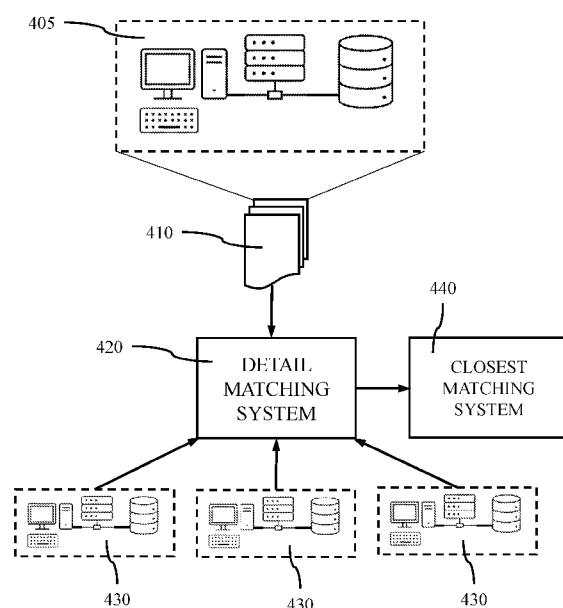


FIG. 4

(57) Abstract: Systems and methods are provided for detail matching. The method includes training a feature classifier (200) to identify technical features, and training a neural network model for a trained importance calculator (300) to calculate an importance value for each identified technical feature. The method further includes receiving a specification sheet (110) including a plurality of technical features, and receiving a plurality of descriptive sheets (120) each including a plurality of technical features. The method further includes identifying the technical features (130) in the specification sheet and the plurality of descriptive sheets using the trained feature classifier (200), and calculating an importance (140) for each identified technical feature using the trained feature importance calculator (300). The method further includes calculating a matching score (150) between the identified technical features of the specification sheet and the identified technical features of the plurality of descriptive sheets based on the importance of each identified technical feature.

Published:

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

TECHNICAL SPECIFICATION MATCHING

RELATED APPLICATION INFORMATION

[0001] This application claims priority to U.S. Provisional Application No. 63/177,406, filed on April 21, 2021 and U.S Patent Application 17/720,431, filed on April 14, 2022, which is incorporated herein by reference in its entirety.

BACKGROUND

Technical Field

[0002] The present invention relates to using trained language models to match technical specifications and more particularly identifying relevant technical features for hierarchical matching of the technical specifications.

Description of the Related Art

[0003] Word embeddings, learned from massive unstructured text data, are widely-adopted building blocks for natural language processing (NLP), such as document classification, sentence classification, and natural language sequence matching. In the same spirit of learning distributed representations for natural language, many NLP applications also benefit from encoding word sequences (e.g., a sentence or document) into a fixed-length feature vector.

SUMMARY

[0004] According to an aspect of the present invention, a method is provided for detail matching. The method includes training a feature classifier to identify technical features, and training a neural network model for a trained importance calculator to calculate an importance value for each identified technical feature. The method further

includes receiving a specification sheet including a plurality of technical features, and receiving a plurality of descriptive sheets each including a plurality of technical features. The method further includes identifying the technical features in the specification sheet and the plurality of descriptive sheets using the trained feature classifier, and calculating an importance for each identified technical feature using the trained feature importance calculator. The method further includes calculating a matching score between the identified technical features of the specification sheet and the identified technical features of the plurality of descriptive sheets based on the importance of each identified technical feature.

[0005] According to another aspect of the present invention, a computer system is provided for detail matching. The system includes one or more processors, a computer memory in electronic communication with the one or more processors, and a display screen in electronic communication with the computer memory and the one or more processors, wherein the computer memory includes a feature classifier trained to identify technical features, a neural network model configured as a trained importance calculator for calculating an importance value for each identified technical feature, text data including a specification sheet including a plurality of technical features, and a plurality of descriptive sheets each including a plurality of technical features, wherein the trained feature classifier identifies the technical features in the specification sheet and the plurality of descriptive sheets, a feature importance calculator to calculate an importance for each identified technical feature using the trained feature importance calculator, and a feature matching system to calculate a matching score between the identified technical features of the specification sheet and the identified technical features of the plurality of descriptive sheets based on the calculated importance of each

identified technical feature, wherein a closest matching product is presented to a user on the display screen.

[0006] According to another aspect of the present invention, a non-transitory computer readable storage medium comprising a computer readable program for detail matching is provided. The computer readable program when executed on a computer causes the computer to perform the steps of training a feature classifier to identify technical features, and training a neural network model for a trained importance calculator to calculate an importance value for each identified technical feature. The computer readable program when executed on a computer also causes the computer to perform the steps of receiving a specification sheet including a plurality of technical features, receiving a plurality of descriptive sheets each including a plurality of technical features, and identifying the technical features in the specification sheet and the plurality of descriptive sheets using the trained feature classifier. The computer readable program when executed on a computer also causes the computer to perform the steps of calculating an importance for each identified technical feature using the trained feature importance calculator, and calculating a matching score between the identified technical features of the specification sheet and the identified technical features of the plurality of descriptive sheets based on the importance of each identified technical feature.

[0007] These and other features and advantages will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF DRAWINGS

[0008] The disclosure will provide details in the following description of preferred embodiments with reference to the following figures wherein:

[0009] FIG. 1 is a block/flow diagram illustrating a high-level system/method for matching technical user needs to manufacturer/developer descriptions is illustratively depicted in accordance with one embodiment of the present invention;

[0010] FIG. 2 is a block/flow diagram illustrating a system/method for a feature/entity classifier, in accordance with an embodiment of the present invention;

[0011] FIG. 3 is a block/flow diagram illustrating a system/method for an entity importance recognizer, in accordance with an embodiment of the present invention;

[0012] FIG. 4 is a block/flow diagram illustrating a process of receiving specification sheet(s) provided by a user/consumer and descriptive material from vendors/manufacturers/developers to identify computer systems/software that best meets the buyer's technical specifications, in accordance with an embodiment of the present invention; and

[0013] FIG. 5 illustrates a computer system for detail matching, in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0014] In accordance with embodiments of the present invention, systems and methods are provided for matching technical specifications of hardware and software of consumers with the technical specifications of producers' hardware and software descriptions. Many times end users desire particular capacities of hardware systems and software applications, but the descriptions used by the manufactures and software developers do not directly coincide with the detailed language used by the prospective

consumer. The technical descriptions provided by vendors also can be complex and use technical jargon that is not easily relatable, requiring technically trained people to review the descriptive materials. Searching for appropriate systems and software can be made more efficient by utilizing trained artificial intelligence to pool descriptive materials and identify hardware and/or software that meets the customer's needs. The hardware and software also may not have the customer's desired capabilities, so determinations of which products come closest in multiple categories may be analyzed and determined.

[0015] In addition, the technical specifications published by the manufacturers and developers are not always available in a single document or from a single source, and the material provided by multiple sources is scattered across electronic and print publications, as well as the vendors' websites.

[0016] In various embodiments, multiple sources of hardware/software technical descriptions from suppliers, manufacturers, and/or application developers are pooled and analyzed using natural language programming (NLP) that can extract technical information from a description of technical features intended for procurement and from descriptive materials provided by multiple manufactures and/or suppliers. A matching set of technical details can be generated even when the language used in the descriptive materials and the procurement materials are different.

[0017] In various embodiments, the entities that are relevant to the business need expression of the consumer also can be extracted, and analysis can be focused on such entities. Different importance can be assigned to different entities to emphasize the importance and applicability of particular ones.

[0018] Referring now in detail to the figures in which like numerals represent the same or similar elements and initially to FIG. 1, a high-level system/method for

matching technical user needs to manufacturer/developer descriptions is illustratively depicted in accordance with one embodiment of the present invention.

[0019] In one or more embodiments, a matching process 100 can involve collecting a user's/consumer's needs document, for example, a request for pricing/proposal (RFP), and vendor description literature, for example, sales brochures, web pages, technical specification sheets and white papers, where the collected documents and literature include text data. The features and details can be extracted from the pooled descriptive materials, including the documents, web pages, and literature (i.e., text data). Pairs of the technical need document (i.e., specification sheet) and the vendor descriptions can be generated, and a matching score calculated for each pair of specification sheet and vendor description. The highest scoring pair(s) can be provided to the user/consumer for consideration for procurement.

[0020] At block 110, a specification sheet(s) provided by a user/consumer can be received and inputted into an extractor. The specification sheet may describe desired features, properties, capabilities, and characteristics of a product in specific terminology and language, that may be different from a producer's description. The specification sheet may describe desired properties, such as processor cores, processor speed, amount of memory, floating point calculation rates, memory transfer rates, data communication rates, communication port capacities, network security properties, etc.

[0021] At block 120, collected descriptive material from vendors/manufacturers/developers can be pooled and inputted into the extractor. The pooled descriptive material can be associated with the vendor/manufacturer/developer entity that supplied the pooled descriptive material.

[0022] At block 130, features and details (collectively referred to as features) can be extracted from each of the specification sheet(s) and the pooled descriptive materials.

The extracted features and details can include the technical specifications, as well as the business entities that provided the specification sheet(s) and the pooled descriptive materials. The technical features/details and business entities can be extracted by a trained classifier. Each document can be segmented into sentences, and the features extracted sentence by sentence using the trained classifier, where the classifier can be trained using external data unrelated to the specification sheets and descriptive materials.

[0023] At block 140, an importance value can be assigned to each of the technical details and extracted entities, where the importance value can be a weight assigned to each technical detail and extracted entity for use in a matching score calculation. The weights can be assigned by a trained neural network importance model.

[0024] At block 150, a matching score can be calculated for each of the descriptive materials in relation to the specification sheet(s). The matching score can be based on each pair formed by the specification sheet(s) and one of the descriptive materials. The matching score can be calculated based on the weighted sum of feature similarities from different features and entities extracted from the specification sheet(s) and the descriptive materials.

[0025] At block 160, a ranked list of the descriptive materials can be provided to a user to identify the system/application that best meets the needs of the user/consumer, where the entity associated with each of the descriptive materials can also be identified and provided. The user/consumer can use the ranked list to identify a vendor entity and system/application that best meets their needs, and initiate a procurement process with the related business entity.

[0026] Referring now to FIG. 2, a block/flow diagram of a system/method for a feature/entity classifier, in accordance with an embodiment of the present invention.

[0027] In one or more embodiments, a feature/detail classifier 200 can be trained to identify the technical features and business entity related to each particular item of descriptive material (e.g., sales brochures, web pages, technical specification sheets and white papers) that may be collected as text data. Labels may not be available for the technical features and business entities identified in the specification sheet(s) provided by a user/consumer or the descriptive materials provided by the vendors. External data can be used for data augmentation, and the knowledge transferred from other domains to the need document (specification sheet) domain.

[0028] At block 210, the text data of the descriptive material can be collected, where the descriptive material can be collected by requesting materials from vendors, downloading documents from websites, web page scraping, magazine/trade paper scanning and optical character recognition, as well as other gathering methods. The collected descriptive material can be pooled in a database for later retrieval and use.

[0029] At block 215, the text data of the descriptive material can be prepared by segregating the text of the descriptive materials into sentences and noun phrases. Each document (e.g., specification sheet, descriptive sheet, web page, etc.) can be segmented into sentences, and the details/features can be extracted sentence by sentence using a trained feature classifier.

[0030] At block 218, noun phrases can be extracted from the text data and the noun phrases identified as relating to technical features and details or business entities. The business entities can be identified using a named entity recognition (NER) tool. The technical features and details can be identified using a trained natural language processing (NLP) model.

[0031] At block 220, a vector representation of the sentences and/or noun phrases for the identified entity and technical details/features can be generated using a trained

Bidirectional Encoder Representations from Transformers (BERT) model. In various embodiments, the vector representation can be generated by a post-trained BERT model to generate the word embeddings for positive and unlabeled entities and features, where positive means the need document/descriptive material pair is a match. A teacher model f_{tea} (e.g., a pre-trained language model) can be used to assign pseudo-labels to unlabeled data that is used to train a student model f_{stu} . Given a pre-trained language model (e.g., BERT) as the teacher model, we first use the distant labels from source domain to fine-tune it to make it adapt to the source domain. The fine-tuned teacher model can be used to generate pseudo labels for the large amount of unlabeled data from both source and target domains. The teacher and student model can exchange knowledge and the training schedules are repeated till convergence. Few-shot labeled data can be utilized in target domain to help select high quality pseudo labels.

[0032] At block 225, the BERT model can be post-trained on business-related documents to make BERT better represent the business entities and technical details.

[0033] At block 230, the vector representations for the features/entities extracted using the post-trained BERT model can be generated. Given a feature/entity, first it is tokenized by a Bidirectional Encoder Representations from Transformers (BERT) tokenizer to generate tokens, then the tokens are input together into the BERT model to generate embeddings for each token, the entity embedding can be generated by summing the embeddings of these tokens to obtain the final embedding of entity, which can be, for example, a 768 dimension vector. A vector representation $v_x = \text{BERT}(x)$ for text, x , which can be a sentence or noun phrase, can be generated. The BERT model can be post-trained on business/technical documents to make BERT better represent the features/details/entities in an application scenario.

[0034] At block 240, the positive unlabeled (PU) entities and features can be used to train a classifier to identify the technical details and business entities in the descriptive materials. A positive-unlabeled (PU) learning method can be used to train the classifier.

[0035] The training process of the feature classifier is described as follows. Given a positive entity set, P , and an unlabeled entity set U , and $E = P \cup U$, where E is the whole feature (entity) set. In addition, there can be a vector representation v_e of $e \in E$, where E is the entire feature set.

[0036] First: Fit a classifier (e.g., Random Forest) to predict the probability that a given feature/entity $e \in E$ is labeled, $p(s=1|v_e)$.

[0037] Second: Use the classifier in Step 1 to predict the probability that the positive features are labeled, $p(s=1|y=1|v_e)$, $e \in P$. The mean of these predicted probabilities will be $p(s=1|y=1)$.

[0038] Third: Use the classifier in Step 1 to estimate the probability that noun phrase, x , $x \in E$ is labeled $p(s=1|v_x)$.

[0039] Fourth: By dividing $p(s=1|v_x)$ in Step 3 by $p(s=1|y=1)$ in Step 2, we can get the probability the feature in x is a technical detail or entity.

[0040] At block 250, the labels and embedding vectors can be input into the trained classifier for training. Because there may be no labels for technical details and business entities, external data can be utilized to learn a feature classifier, $F(v_x)$, and transfer it to a business need domain. For each noun phrase, x , its vector representation v_x (e.g., generated using a Bidirectional Encoder Representations from Transformers (BERT) model) is fed into the classifier, and the classifier judges whether the input noun phrase is a technical detail or business entity. The features that indicate company needs can be extracted from a specification sheet to represent a need document, where $doc=(e_1, e_2, \dots, e_m)$.

[0041] At block 260, noun phrases are extracted from the descriptive materials. For each noun phrase, x , whether it exists in a positive entity set can be checked.

[0042] Specifically, text data is first collected from national custom goods category and an industry category. With noun phrases extracted from them by existing tools, the extracted noun phrases can be treated as business entities and labeled as positive. In addition, the webpages of companies who list their needs can be collected, and an online named entity recognition (NER) tool can be used to identify the entities in the Consumer Good type as business entities and label them as positive. These business entities form a positive biz-entity set.

[0043] FIG. 3 is a block/flow diagram of a system/method for an entity importance recognizer is illustratively depicted in accordance with an embodiment of the present invention.

[0044] A procedure of feature/entity importance learning 300 can learn to identify the technical details of greatest importance to the user/consumer by extracting the features from a specification/needs document supplied by the user/consumer.

[0045] At block 310, triplets including the user's/consumer's needs document and the descriptive materials can be generated, where the descriptive materials can include a positive document and a negative document. For a given need document doc_i , there are M_i positive need documents and N_i negative need documents, where positive means the need document pair is a match, and negative means the pair is not a match.

[0046] In various embodiments, there can be generated, $M_i * N_i$ triplet (doc_i, doc_p, doc_q) for doc_i , where doc_p is a need document from the M_i positives, and doc_q is a need document from the N_i negatives. If there are K need documents (specification sheets), then in total we can generate:

[0047] $T = \sum_{i=1}^k M_i * N_i$ triplets, represented as T .

[0048] At block 320, the vector representations for the extracted features and entities can be generated using the post-trained BERT model. Given a feature/entity, first it is tokenized by a BERT tokenizer to generate tokens, then the tokens are input together into the BERT model to generate embeddings for each token, the entity embedding can be generated by summing the embeddings of these tokens to obtain the final embedding of entity, which can be, for example, a 768 dimension vector.

[0049] At block 330, entity importance is learned.

[0050] The importance of feature/entity, e , can be calculated by model $H(v_e)$. The above equation looks for its best match in E_c . To make the w_e flexible and easy extend to unseen entities, we further model it as $w_e = H(v_e) = v_e * \theta + b$, where θ and b are two parameters to be learned.

[0051] Given a triplet $t(\text{doc}_i, \text{doc}_p, \text{doc}_q)$, where doc_p is a matched document of doc_i , and doc_q is a unmatched document of doc_i , we aim to learn the entity importance to maximize the matching score (calculated as above) between doc_i and doc_p , meanwhile minimizing the matching score between doc_i and doc_q . Given a triplet $t(\text{doc}_i, \text{doc}_p, \text{doc}_q)$, the parameters of $H(v_e)$ can be tuned to maximize the matching score between doc_i and doc_p , meanwhile minimize the matching score between doc_i and doc_q .

[0052] At block 340, a loss function can be formulated.

[0053] For each triplet, the loss function can be described as follows:

$$[0054] \quad L(t) = \max(0, (1 - s_{i,p}) - (1 - s_{i,q}) + \alpha)$$

[0055] where α is a margin between positive and negative pairs.

[0056] Further with all the triplets in T , we can formulate the cost function, that is the sum of all losses, to be used for minimization of the following optimization problem. In this way, we can learn the entity importance for each entity, which later can be used for the matching score calculation.

[0057] $L = \sum_{t \in T} L(t)$

[0058] At block 350, a matching score can be calculated. Let E_q and E_c be entities appearing in Doc_q and Doc_c , respectively. The matching score, s , is evaluated by the following equation,

[0059] $s_{q,c} = \sum_{e_q \in E_q} w_{e_q} \max_{e_c \in E_c} \frac{v_{e_q} \cdot v_{e_c}}{\|v_{e_q}\| \|v_{e_c}\|},$

[0060] where v_e denotes the vector semantic representation (e.g., word2vec) for feature/entity e , and w_e is the importance for feature/entity, e .

[0061] For each pair of query and candidate documents, a feature-based matching score is used to evaluate the similarity of two need documents. A matching score can be calculated for each pair of the user's/consumer's specification document and the descriptive materials for a particular product from a vendor/manufacture/developer.

[0062] FIG. 4 is a block/flow diagram illustrating a process of receiving specification sheet(s) provided by a user/consumer and descriptive material from vendors/manufacturers/developers to identify computer systems/software that best meets the buyer's technical specifications, in accordance with an embodiment of the present invention.

[0063] In one or more embodiments, a detail matcher system 420 can execute a matching process 100 that identifies the best suited vendor product 430 from the descriptive materials available for the products 430 and outputs the closest matching system 440.

[0064] In block 410, the specification sheet(s) including the desired technical features/details provided by a prospective buyer can be inputted into the detail matching system 420. The various descriptive materials obtained for each of a plurality of vendor systems can be inputted into the detail matching system 420. The detail matching system 420 can perform a matching process 100 utilizing trained neural networks that

extract the details from the specification sheet(s) and the descriptive materials available for the products 430. An importance model $w_e = H(v_e)$ can be used to identify and calculate the importance of the various details, for example, to identify whether memory size, processor speed, number of cores, or bus speed is most important in relation to the available vendor products 430. A procurement order can be generated for the closest matching system 440 identified by the detail matching system 420.

[0065] FIG. 5 illustrates a computer system for detail matching, in accordance with an embodiment of the present invention.

[0066] In one or more embodiments, the computer matching system 500 for detail/feature matching of technical features of computer systems and/or software can include one or more processors 510, which can be central processing units (CPUs), graphics processing units (GPUs), and combinations thereof, and a computer memory 520 in electronic communication with the one or more processors 510, where the computer memory 520 can be random access memory (RAM), solid state drives (SSDs), hard disk drives (HDDs), optical disk drives (ODD), etc. The memory 520 can be configured to store the detail matching tool 420, including a trained feature classifier 200, trained feature/entity importance calculator 300, training corpus 550, and collected text data 210. The feature classifier 200 can be a neural network configured to identify technical features and details in a specification sheet and in the collected text data 210. The feature/entity importance calculator 300 can be a neural network configured to determine the relative importance of each of the identified features/details in the provided specification sheets and descriptive materials. The training corpus 550 can be used to train the different neural networks, where the training corpus can contain external data. A display module can be configured to present an ordered list of the vendor products that match the customer's specification sheet(s) to a user on a display

screen 530, as a summary of the available systems/software. The memory 520 and one or more processors 510 can be in electronic communication with a display screen 530 over a system bus and I/O controllers, where the display screen 530 can present the ranked list of available systems/software and/or prepare a procurement order for the highest ranked available system/software.

[0067] Embodiments described herein may be entirely hardware, entirely software or including both hardware and software elements. In a preferred embodiment, the present invention is implemented in software, which includes but is not limited to firmware, resident software, microcode, etc.

[0068] Embodiments may include a computer program product accessible from a computer-usable or computer-readable medium providing program code for use by or in connection with a computer or any instruction execution system. A computer-usable or computer readable medium may include any apparatus that stores, communicates, propagates, or transports the program for use by or in connection with the instruction execution system, apparatus, or device. The medium can be magnetic, optical, electronic, electromagnetic, infrared, or semiconductor system (or apparatus or device) or a propagation medium. The medium may include a computer-readable storage medium such as a semiconductor or solid state memory, magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disk and an optical disk, etc.

[0069] Each computer program may be tangibly stored in a machine-readable storage media or device (e.g., program memory or magnetic disk) readable by a general or special purpose programmable computer, for configuring and controlling operation of a computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system may also be considered to be

embodied in a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions described herein.

[0070] A data processing system suitable for storing and/or executing program code may include at least one processor coupled directly or indirectly to memory elements through a system bus. The memory elements can include local memory employed during actual execution of the program code, bulk storage, and cache memories which provide temporary storage of at least some program code to reduce the number of times code is retrieved from bulk storage during execution. Input/output or I/O devices (including but not limited to keyboards, displays, pointing devices, etc.) may be coupled to the system either directly or through intervening I/O controllers.

[0071] Network adapters may also be coupled to the system to enable the data processing system to become coupled to other data processing systems or remote printers or storage devices through intervening private or public networks. Modems, cable modem and Ethernet cards are just a few of the currently available types of network adapters.

[0072] As employed herein, the term “hardware processor subsystem” or “hardware processor” can refer to a processor, memory, software or combinations thereof that cooperate to perform one or more specific tasks. In useful embodiments, the hardware processor subsystem can include one or more data processing elements (e.g., logic circuits, processing circuits, instruction execution devices, etc.). The one or more data processing elements can be included in a central processing unit, a graphics processing unit, and/or a separate processor- or computing element-based controller (e.g., logic gates, etc.). The hardware processor subsystem can include one or more on-board memories (e.g., caches, dedicated memory arrays, read only memory, etc.). In some

embodiments, the hardware processor subsystem can include one or more memories that can be on or off board or that can be dedicated for use by the hardware processor subsystem (e.g., ROM, RAM, basic input/output system (BIOS), etc.).

[0073] In some embodiments, the hardware processor subsystem can include and execute one or more software elements. The one or more software elements can include an operating system and/or one or more applications and/or specific code to achieve a specified result.

[0074] In other embodiments, the hardware processor subsystem can include dedicated, specialized circuitry that performs one or more electronic processing functions to achieve a specified result. Such circuitry can include one or more application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), and/or programmable logic arrays (PLAs).

[0075] These and other variations of a hardware processor subsystem are also contemplated in accordance with embodiments of the present invention.

[0076] Reference in the specification to “one embodiment” or “an embodiment” of the present invention, as well as other variations thereof, means that a particular feature, structure, characteristic, and so forth described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, the appearances of the phrase “in one embodiment” or “in an embodiment”, as well as any other variations, appearing in various places throughout the specification are not necessarily all referring to the same embodiment. However, it is to be appreciated that features of one or more embodiments can be combined given the teachings of the present invention provided herein.

[0077] It is to be appreciated that the use of any of the following “/”, “and/or”, and “at least one of”, for example, in the cases of “A/B”, “A and/or B” and “at least one of

A and B”, is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of both options (A and B). As a further example, in the cases of “A, B, and/or C” and “at least one of A, B, and C”, such phrasing is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of the third listed option (C) only, or the selection of the first and the second listed options (A and B) only, or the selection of the first and third listed options (A and C) only, or the selection of the second and third listed options (B and C) only, or the selection of all three options (A and B and C). This may be extended for as many items listed.

[0078] The foregoing is to be understood as being in every respect illustrative and exemplary, but not restrictive, and the scope of the invention disclosed herein is not to be determined from the Detailed Description, but rather from the claims as interpreted according to the full breadth permitted by the patent laws. It is to be understood that the embodiments shown and described herein are only illustrative of the present invention and that those skilled in the art may implement various modifications without departing from the scope and spirit of the invention. Those skilled in the art could implement various other feature combinations without departing from the scope and spirit of the invention. Having thus described aspects of the invention, with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.

WHAT IS CLAIMED IS:

1. A method of detail matching (100), comprising:
training a feature classifier (200) to identify technical features;
training a neural network model for a trained importance calculator (300) to calculate an importance value for each identified technical feature;
receiving a specification sheet (110) including a plurality of technical features;
receiving a plurality of descriptive sheets (120) each including a plurality of technical features;
identifying the technical features (130) in the specification sheet and the plurality of descriptive sheets using the trained feature classifier (200);
calculating an importance for each identified technical feature (140) using the trained feature importance calculator (300); and
calculating a matching score (150) between the identified technical features of the specification sheet and the identified technical features of the plurality of descriptive sheets based on the importance of each identified technical feature.
2. The method of claim 1, wherein the trained importance calculator is trained using triplets of the specification sheet and the plurality of descriptive sheets.
3. The method of claim 2, further comprising generating vector embeddings for each identified technical feature using a trained Bidirectional Encoder Representations from Transformers (BERT) model.

4. The method of claim 3, wherein the matching scores, $s_{q,c}$, are calculated using $s_{q,c} = \sum_{e_q \in E_q} w_{e_q} \max_{e_c \in E_c} \frac{v_{e_q} \cdot v_{e_c}}{\|v_{e_q}\| \|v_{e_c}\|}$, wherein v_e denotes a vector semantic representation for each feature/entity e , and w_e is the importance for each feature/entity, e .

5. The method of claim 4, wherein training the feature classifier utilizes a positive feature set, P , and an unlabeled feature set, U , where $E = P \cup U$, where E is the whole feature set.

6. The method of claim 4, wherein matched documents are utilized to train the entity importance model $H(v_e) = w_e$, where v_e is the vector representation of feature, e , and w_e is the learned feature importance.

7. The method of claim 6, wherein the parameters of the entity importance model $H(v_e) = w_e$, are tuned based on a loss function, $L(t) = \max(0, (1 - s_{i,p}) - (1 - s_{i,q}) + \alpha)$.

8. A computer system for detail matching, comprising:
 one or more processors (510);
 a computer memory (520) in electronic communication with the one or more processors (510); and
 a display screen (530) in electronic communication with the computer memory and the one or more processors;
 wherein the computer memory (520) includes:
 a feature classifier (200) trained to identify technical features;

a neural network model configured as a trained importance calculator (300) for calculating an importance value (140) for each identified technical feature;

text data (210) including a specification sheet including a plurality of technical features, and a plurality of descriptive sheets each including a plurality of technical features, wherein the trained feature classifier (200) identifies the technical features in the specification sheet and the plurality of descriptive sheets; and

a feature matching system (420) to calculate a matching score (150) between the identified technical features of the specification sheet and the identified technical features of the plurality of descriptive sheets based on the calculated importance of each identified technical feature, wherein a closest matching product is presented (160) to a user on the display screen (530).

9. The computer system of claim 8, wherein the trained importance calculator is trained using triplets of the specification sheet and the plurality of descriptive sheets.

10. The computer system of claim 9, wherein feature classifier generates vector embeddings for each identified technical feature using a trained Bidirectional Encoder Representations from Transformers (BERT) model.

11. The computer system of claim 10, wherein the matching scores, $s_{q,c}$, are calculated using $s_{q,c} = \sum_{e_q \in E_q} w_{e_q} \max_{e_c \in E_c} \frac{v_{e_q} \cdot v_{e_c}}{\|v_{e_q}\| \|v_{e_c}\|}$, wherein v_e denotes a vector semantic representation for each feature/entity e , and w_e is the importance for each feature/entity, e .

12. The computer system of claim 11, wherein training the feature classifier utilizes a positive feature set, P , and an unlabeled feature set, U , where $E = P \cup U$, where E is the whole feature set.

13. The computer system of claim 11, wherein matched documents are utilized to train the entity importance model $H(v_e) = w_e$, where v_e is the vector representation of feature, e , and w_e is the learned feature importance.

14. The computer system of claim 13, wherein the parameters of the entity importance model $H(v_e) = w_e$, are tuned based on a loss function, $L(t) = \max(0, (1 - s_{i,p}) - (1 - s_{i,q}) + \alpha)$.

15. A non-transitory computer readable storage medium comprising a computer readable program for detail matching, wherein the computer readable program when executed on a computer causes the computer to perform the steps of:

training a feature classifier (200) to identify technical features;

training a neural network model for a trained importance calculator (300) to calculate an importance value for each identified technical feature;

receiving a specification sheet (110) including a plurality of technical features;

receiving a plurality of descriptive sheets (120) each including a plurality of technical features;

identifying the technical features (130) in the specification sheet and the plurality of descriptive sheets using the trained feature classifier (200);

calculating an importance (140) for each identified technical feature using the trained feature importance calculator (300); and

calculating a matching score (150) between the identified technical features of the specification sheet and the identified technical features of the plurality of descriptive sheets based on the importance of each identified technical feature.

16. The non-transitory computer readable storage medium comprising a computer readable program of claim 15, wherein the trained importance calculator is trained using triplets of the specification sheet and the plurality of descriptive sheets.

17. The non-transitory computer readable storage medium comprising a computer readable program of claim 16, further comprising generating vector embeddings for each identified technical feature using a trained Bidirectional Encoder Representations from Transformers (BERT) model.

18. The non-transitory computer readable storage medium comprising a computer readable program of claim 17, wherein the matching scores, $s_{q,c}$, are calculated using $s_{q,c} = \sum_{e_q \in E_q} w_{e_q} \max_{e_c \in E_c} \frac{v_{e_q} \cdot v_{e_c}}{\|v_{e_q}\| \|v_{e_c}\|}$, wherein v_e denotes a vector semantic representation for each feature/entity e , and w_e is the importance for each feature/entity, e .

19. The non-transitory computer readable storage medium comprising a computer readable program of claim 18, wherein training the feature classifier utilizes a positive feature set, P, and an unlabeled feature set, U, where $E = P \cup U$, where E is the whole feature set.

20. The non-transitory computer readable storage medium comprising a computer readable program of claim 18, wherein matched documents are utilized to train the entity importance model $H(v_e) = w_e$, where v_e is the vector representation of feature, e , and w_e is the learned feature importance, and the parameters of the entity importance model $H(v_e) = w_e$, are tuned based on a loss function, $L(t) = \max(0, (1 - s_{i,p}) - (1 - s_{i,q}) + \alpha)$.

1/5

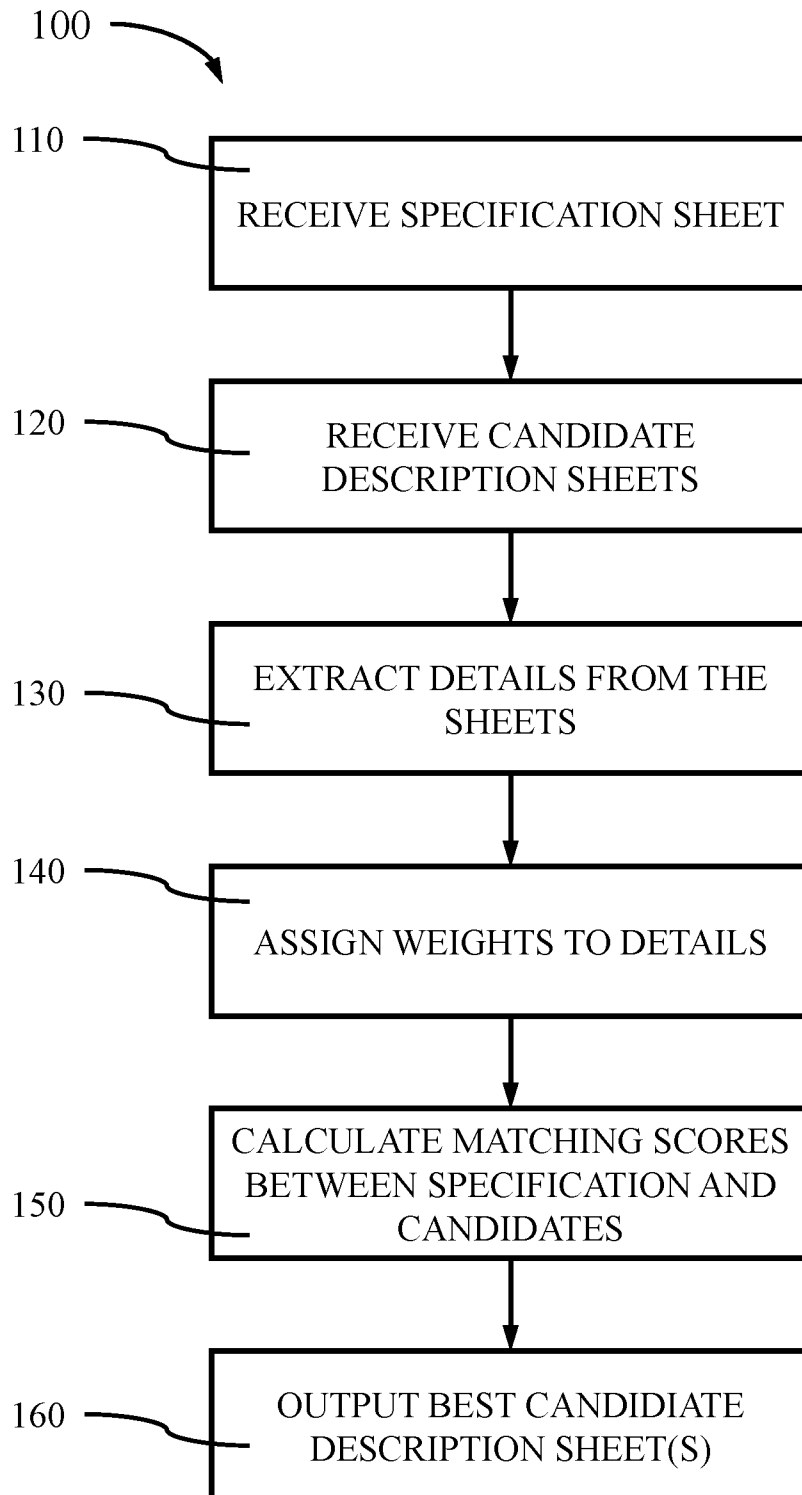


FIG. 1

2/5

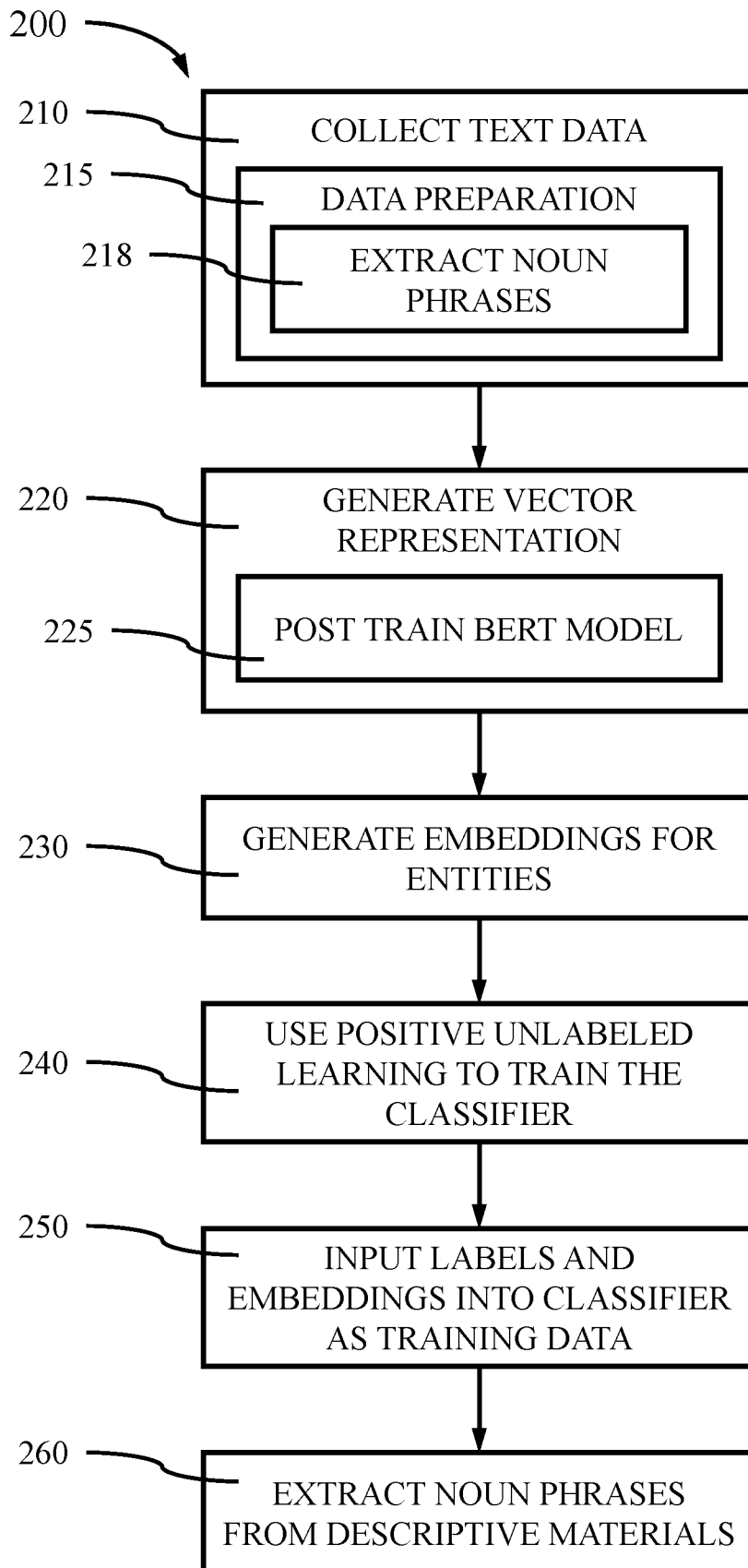


FIG. 2

3/5

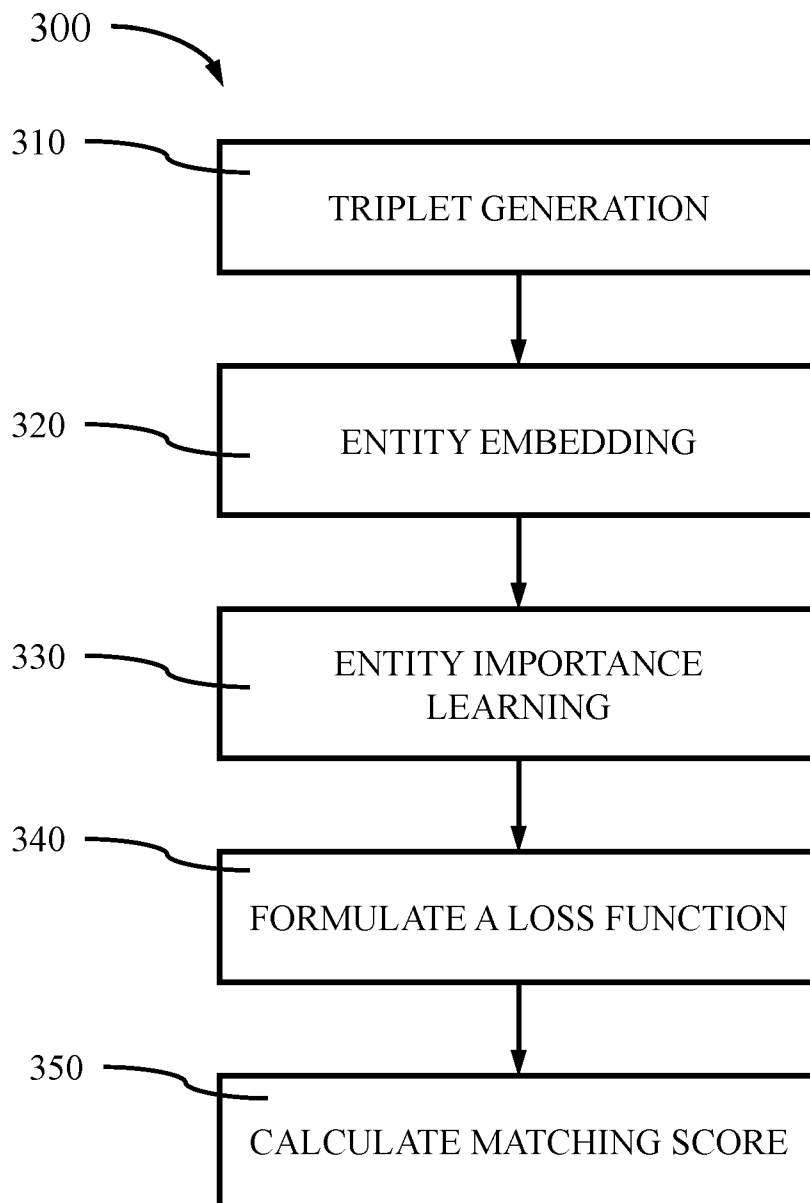


FIG. 3

4/5

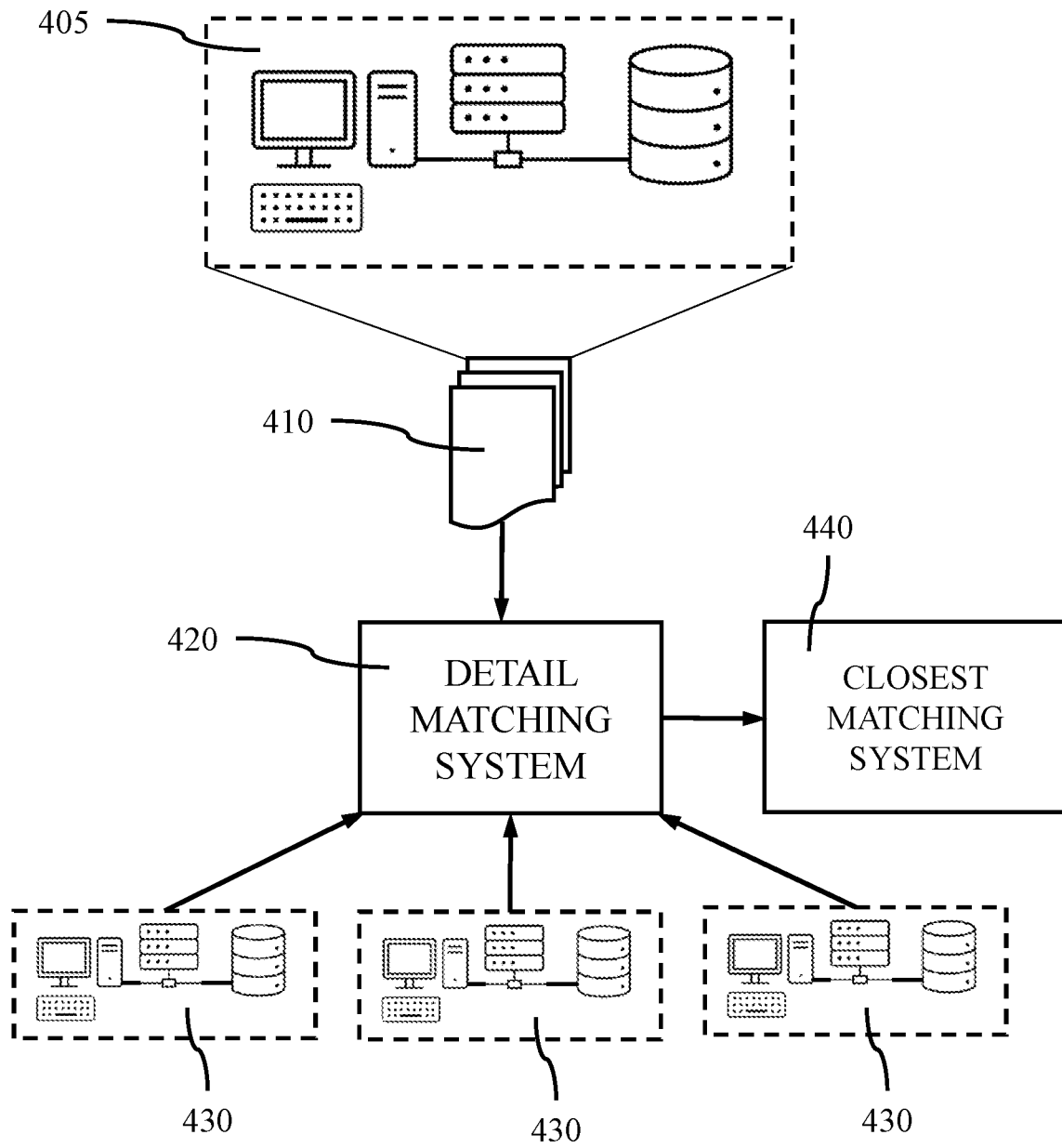


FIG. 4

5/5

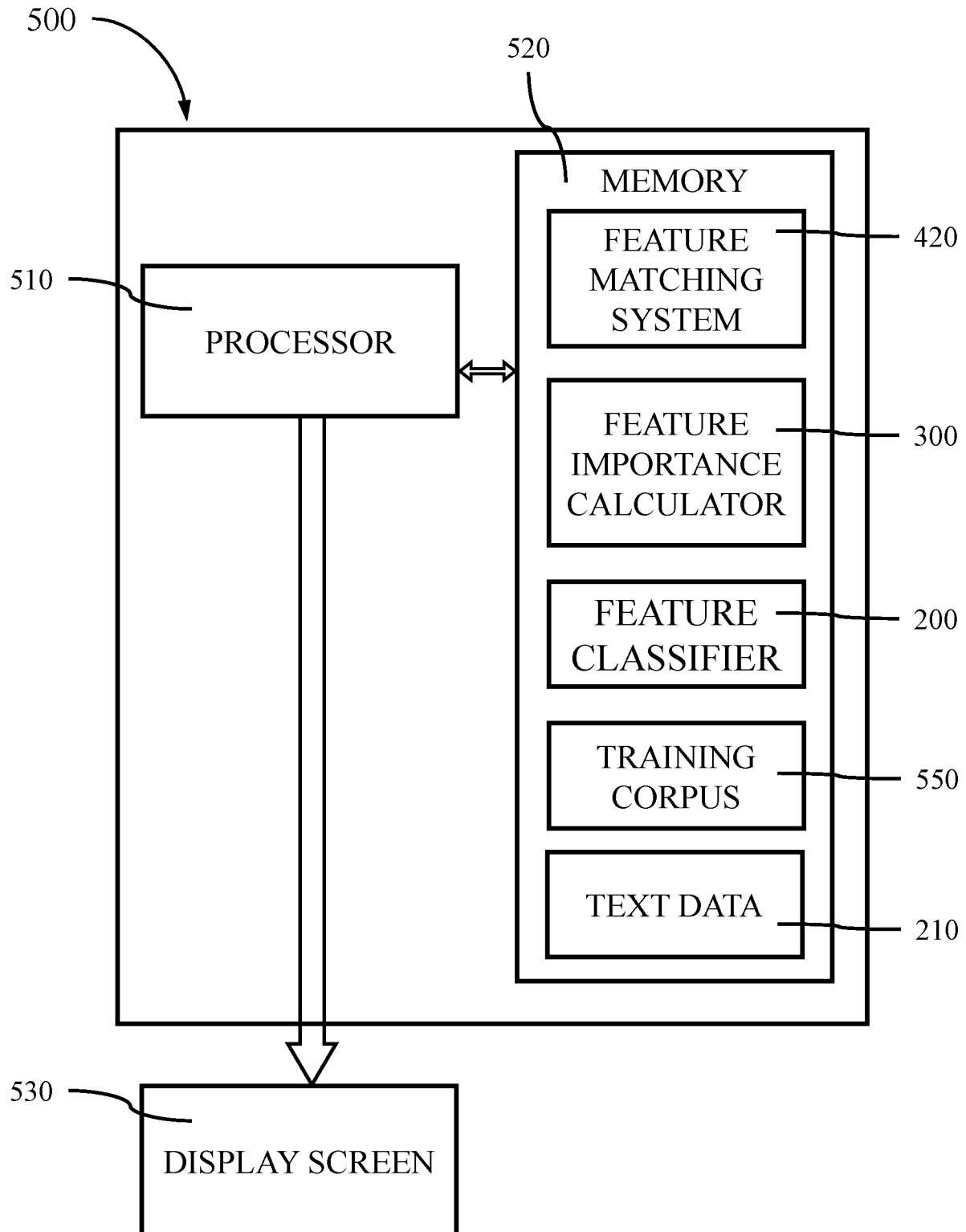


FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2022/024995

A. CLASSIFICATION OF SUBJECT MATTER G06F 16/35(2019.01)i; G06F 40/279(2020.01)i; G06K 9/62(2006.01)i; G06N 3/08(2006.01)i According to International Patent Classification (IPC) or to both national classification and IPC				
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06F 16/35(2019.01); G06F 11/07(2006.01); G06F 15/18(2006.01); G06F 16/33(2019.01); G06F 17/27(2006.01); G06F 17/30(2006.01); G06F 19/00(2011.01) Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Korean utility models and applications for utility models Japanese utility models and applications for utility models Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) eKOMPASS(KIPO internal) & Keywords: feature classifier, technical feature, specification sheet, descriptive sheet, importance, score				
C. DOCUMENTS CONSIDERED TO BE RELEVANT				
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.		
A	US 7194324 B2 (SIEGFRIED SOLLNER) 20 March 2007 (2007-03-20) column 2, line 35; column 3, line 20; column 5, lines 22-23, 40; and claims 1-8	1-20		
A	US 2002-0026297 A1 (FRANK LEYMANN et al.) 28 February 2002 (2002-02-28) paragraphs [0055]-[0084]; claims 1-15; and figure 6	1-20		
A	US 7082426 B2 (TIMOTHY ALLEN MUSGROVE et al.) 25 July 2006 (2006-07-25) column 9, line 13 - column 24, line 53; claims 1-31; and figures 5-9	1-20		
A	CN 110990529 A (I-XINNUO CREDIT CO., LTD. et al.) 10 April 2020 (2020-04-10) paragraphs [0025]-[0063]; claims 1-8; and figures 1-4	1-20		
A	US 2019-0347148 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 14 November 2019 (2019-11-14) paragraphs [0034]-[0079]; claims 1-12; and figures 2-4	1-20		
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.				
<table style="width: 100%; border: none;"> <tr> <td style="width: 50%; vertical-align: top; border: none;"> * 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 </td> <td style="width: 50%; vertical-align: top; border: none;"> "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 </td> </tr> </table>			* 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
* 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 04 August 2022		Date of mailing of the international search report 04 August 2022		
Name and mailing address of the ISA/KR Korean Intellectual Property Office 189 Cheongsa-ro, Seo-gu, Daejeon 35208, Republic of Korea Facsimile No. +82-42-481-8578		Authorized officer BYUN, Sung Cheal Telephone No. +82-42-481-8262		

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US2022/024995

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	7194324	B2	20 March 2007	EP	1383061	A2	21 January 2004
				EP	1383061	A3	23 June 2004
				US	2004-0019397	A1	29 January 2004
US	2002-0026297	A1	28 February 2002	EP	1197849	A2	17 April 2002
				EP	1197849	A3	09 November 2005
				US	6886007	B2	26 April 2005
US	7082426	B2	25 July 2006	EP	1287419	A2	05 March 2003
				EP	1493107	A1	05 January 2005
				EP	1639505	A2	29 March 2006
				EP	1668538	A2	14 June 2006
				JP	2005-522784	A	28 July 2005
				JP	2006-527886	A	07 December 2006
				US	2003-0065643	A1	03 April 2003
				US	2004-0078284	A1	22 April 2004
				US	2004-0143600	A1	22 July 2004
				US	2006-0242192	A1	26 October 2006
				US	2012-0191719	A1	26 July 2012
				US	2012-0233170	A1	13 September 2012
				US	2016-0019611	A1	21 January 2016
				US	2016-0063582	A1	03 March 2016
				US	2016-0086239	A1	24 March 2016
				US	2016-0086240	A1	24 March 2016
				US	6535880	B1	18 March 2003
				US	6714933	B2	30 March 2004
				US	6725222	B1	20 April 2004
				US	8930370	B2	06 January 2015
				US	9177059	B2	03 November 2015
				WO	01-86377	A2	15 November 2001
				WO	01-86377	A3	28 February 2002
				WO	2005-001595	A2	06 January 2005
				WO	2005-001595	A3	09 September 2005
				WO	2005-036306	A2	21 April 2005
CN	110990529	A	10 April 2020	None			
US	2019-0347148	A1	14 November 2019	CN	110472809	A	19 November 2019