US 20210256354A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2021/0256354 A1

CHOI et al. (43) Pub. Date: Aug. 19, 2021

(54) **ARTIFICIAL INTELLIGENCE LEARNING-BASED USER KNOWLEDGE TRACING SYSTEM AND OPERATING METHOD THEREOF**

(71) Applicant: **RIIID INC.**, Seoul (KR)

(72) Inventors: **Young Duck CHOI**, Seoul (KR); **Young Nam LEE**, Seoul (KR); **Jung Hyun CHO**, Seoul (KR); **Jin Eon BAEK**, Seoul (KR); **Byung Soo KIM**, Seoul (KR); **Yeong Min CHA**, Gyeonggido (KR); **Dong Min SHIN**, Seoul (KR); **Chan BAE**, Seoul (KR); **Jae We HEO**, Seoul (KR)

(21) Appl. No.: **17/177,196**

(22) Filed: **Feb. 16, 2021**

(30) **Foreign Application Priority Data**

Feb. 18, 2020 (KR) 10-2020-0019853
May 14, 2020 (KR) 10-2020-0057446

Publication Classification

(51) **Int. Cl.**

G06N 3/04	(2006.01)
G06N 5/02	(2006.01)
G06F 17/18	(2006.01)
G06F 17/16	(2006.01)

(52) **U.S. Cl.**

CPC ***G06N 3/0454*** (2013.01); ***G06F 17/16*** (2013.01); ***G06F 17/18*** (2013.01); ***G06N 5/02*** (2013.01)

(57) **ABSTRACT**

The present invention relates to a user knowledge tracing method with more improved accuracy, and an operating method for a user knowledge tracing system including a plurality of encoder neural networks and a plurality of decoder neural networks includes: inputting exercise information to a k-th encoder neural network and inputting response information to a k-th decoder neural network; generating query data, which is information on an exercise for which a user is to predict a correct answer probability, by reflecting a weight to the response information and generating attention information to be used as a weight for the query data by reflecting the weight to the exercise information; and training the user knowledge tracing system by using the attention information as the weight for the query data.

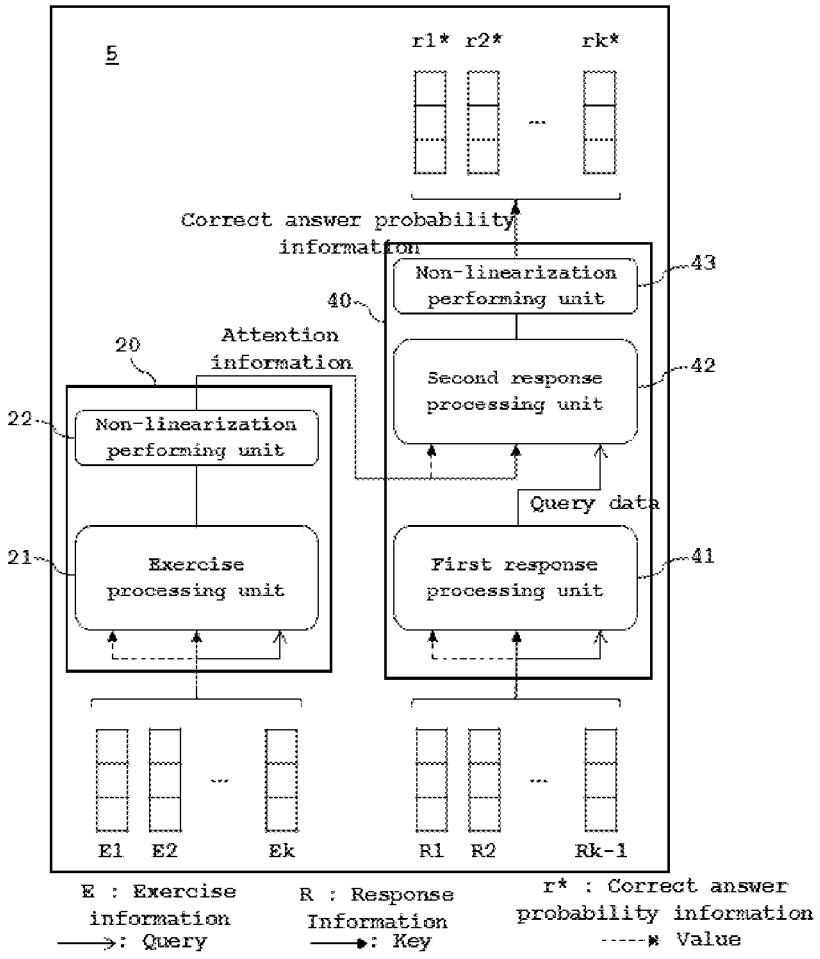


FIG 1

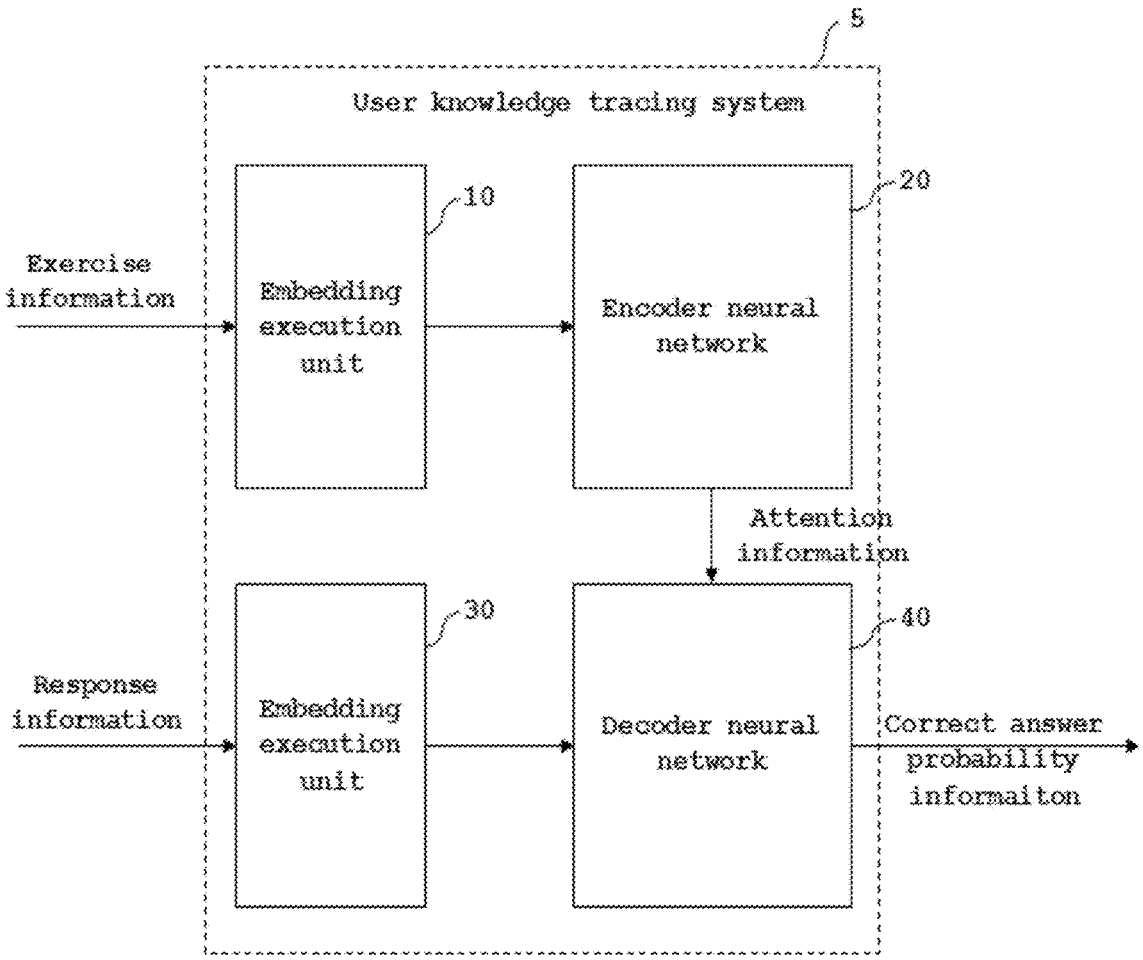

5
User knowledge tracing system
10
20
Exercise information
Embedding execution unit
Encoder neural network
Attention information
30
40
Response information
Embedding execution unit
Decoder neural network
Correct answer probability informaiton

FIG 2

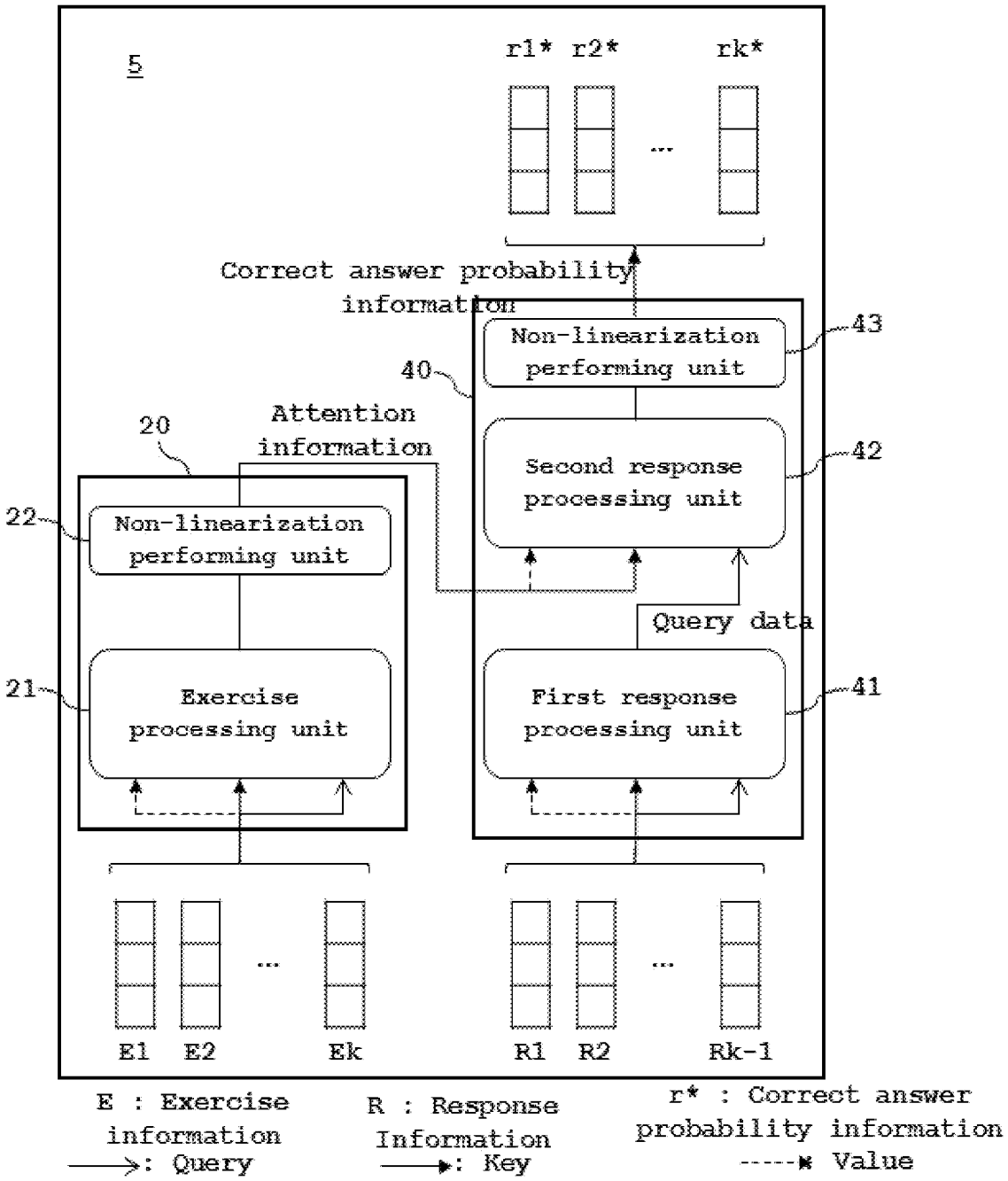
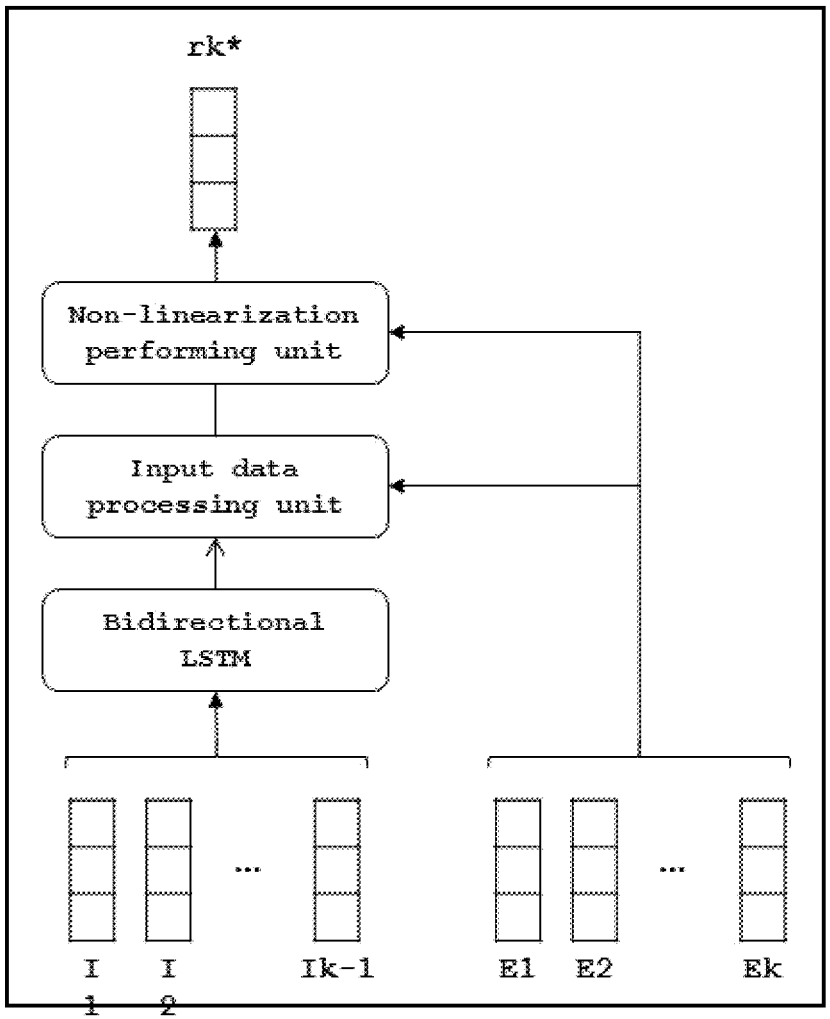

5
r1* r2* rk*
Correct answer probability information
43
Non-linearization performing unit
40
Attention information
20
42
Second response processing unit
22
Non-linearization performing unit
Query data
21
Exercise processing unit
First response processing unit
41
E1 E2 Ek
R1 R2 Rk-1
E : Exercise information
: Query
R : Response Information
: Key
r* : Correct answer probability information
Value

FIG 3



E : Exercise information

I : Interaction information

r* : Correct answer probability information

⟶: Target ⟶ Source

FIG 4

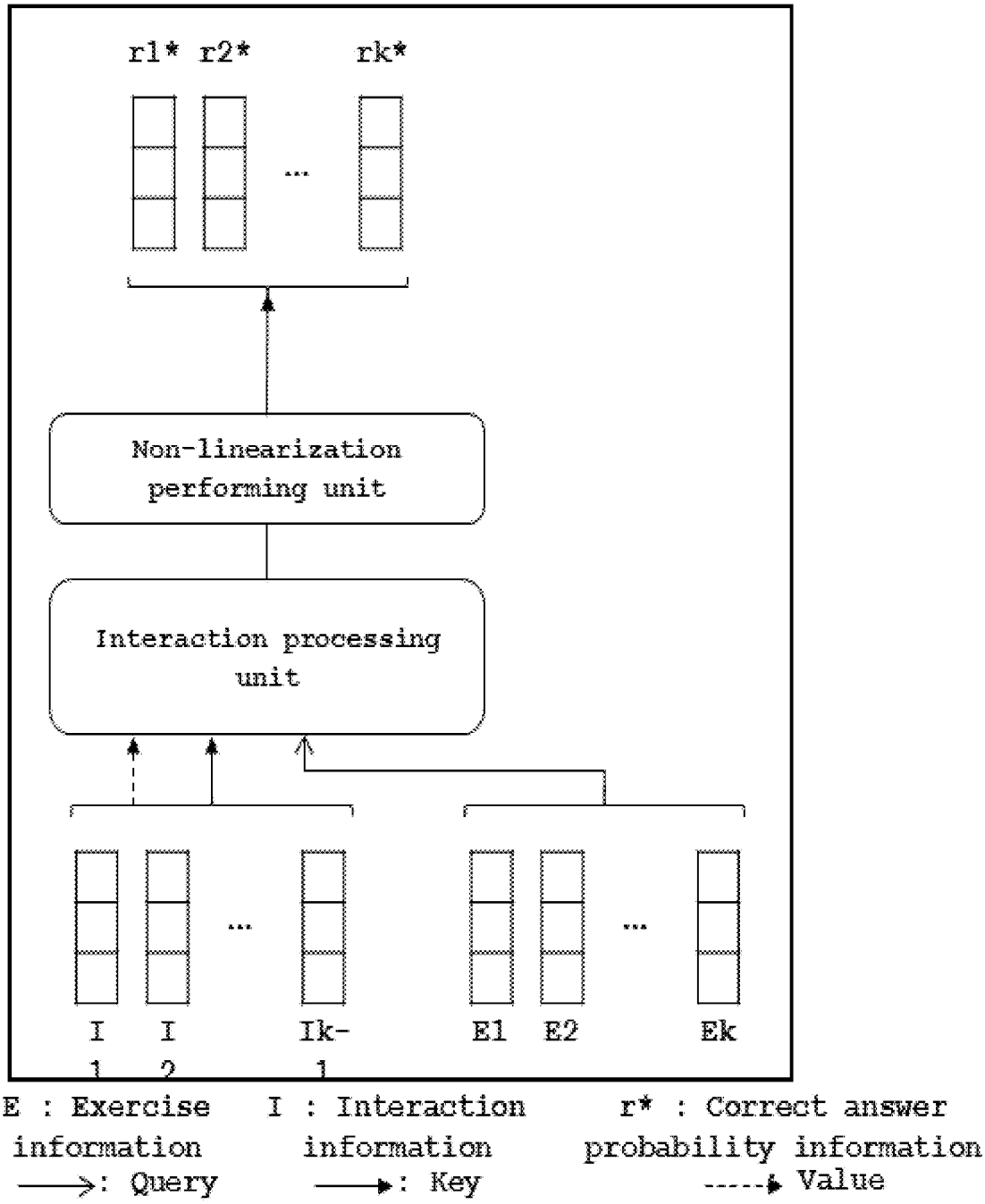

r1* r2* rk*
...
Non-linearization performing unit
Interaction processing unit
...
...
I 1
I 2
Ik-1
E1
E2
Ek
E : Exercise information
I : Interaction information
r* : Correct answer probability information
: Query
: Key
: Value

FIG 5

Input data

	Exercise information (E)	Interaction information (I)	Response information (R)
Exercise identification information	✓	✓	
Exercise category information	✓	✓	
Position information	✓	✓	✓
Response accuracy information		✓	✓
Required time information		✓	
Time recording information		✓	

FIG 6

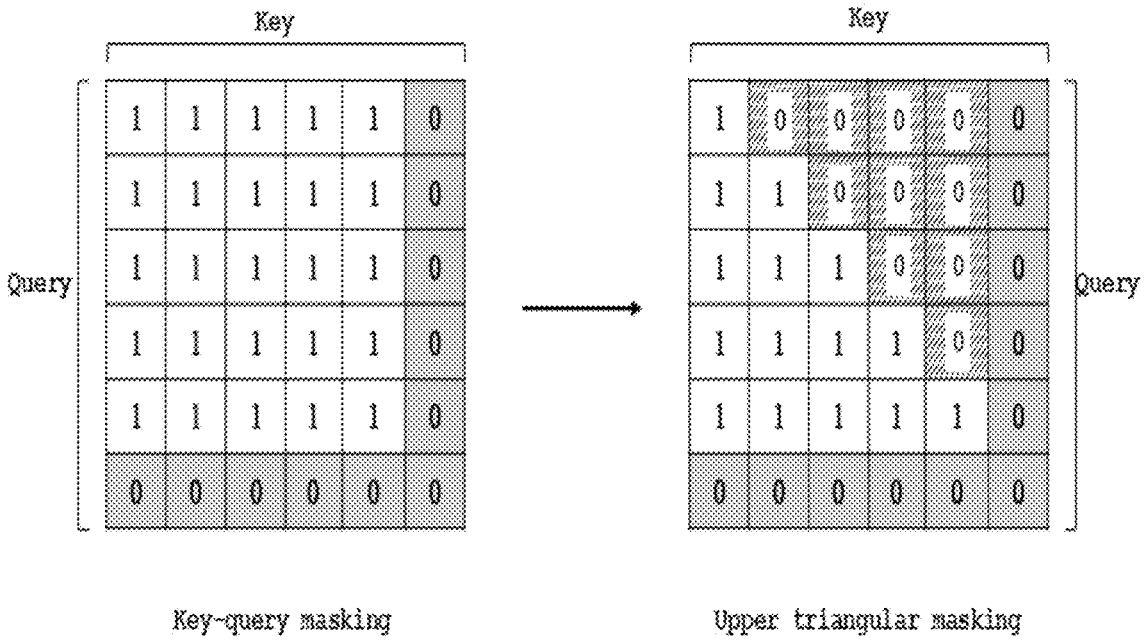

Key
Query
1 1 1 1 1 0
1 1 1 1 1 0
1 1 1 1 1 0
1 1 1 1 1 0
1 1 1 1 1 0
0 0 0 0 0 0
Key
Query
1 0 0 0 0 0
1 1 0 0 0 0
1 1 1 0 0 0
1 1 1 1 0 0
1 1 1 1 1 0
0 0 0 0 0 0
Key-query masking
Upper triangular masking

FIG 7

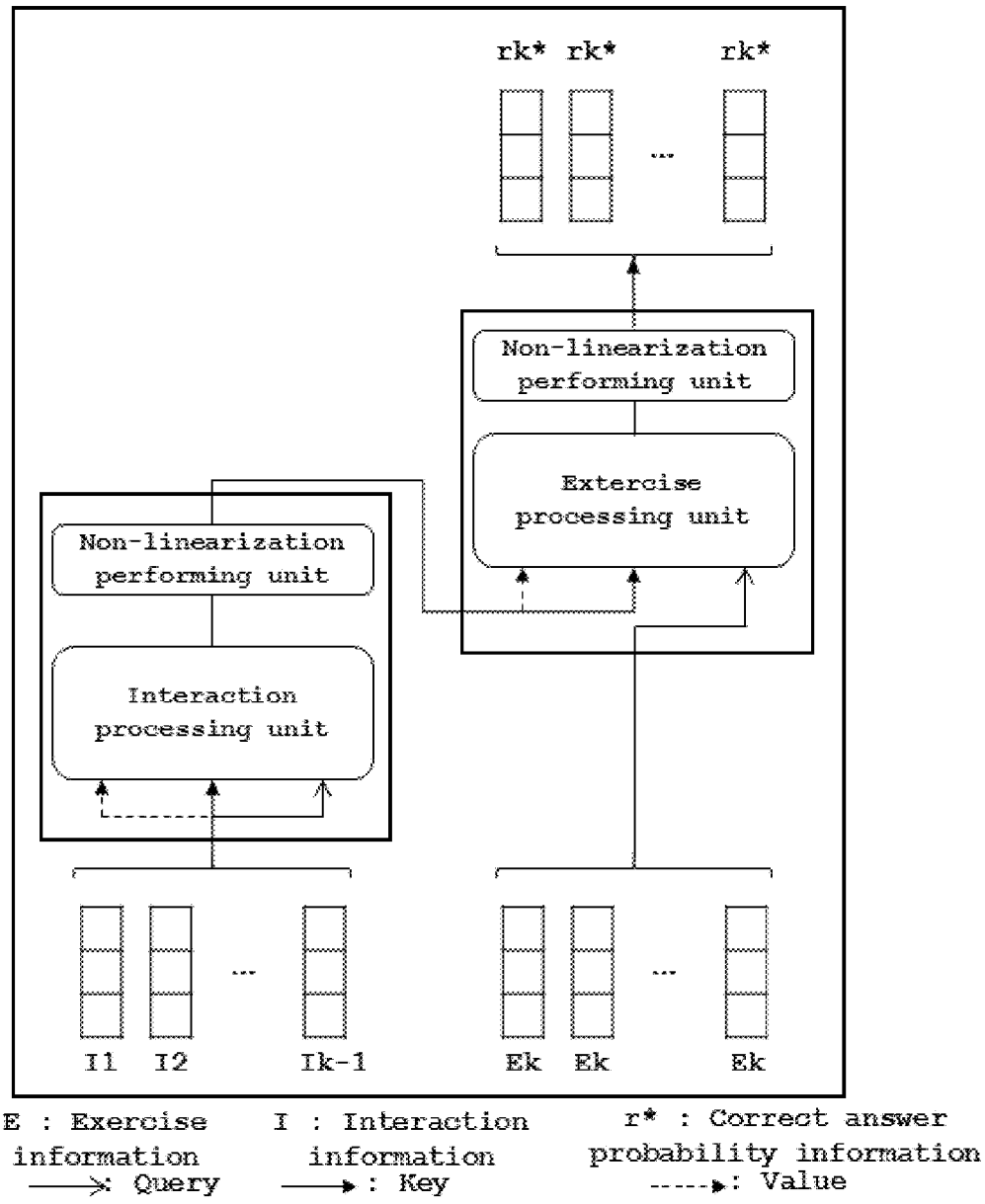

rk* rk* rk*
Non-linearization performing unit
Extercise processing unit
Non-linearization performing unit
Interaction processing unit
I1 I2 Ik-1
Ek Ek Ek
E : Exercise information
Query
I : Interaction information
Key
r* : Correct answer probability information
Value

FIG 8

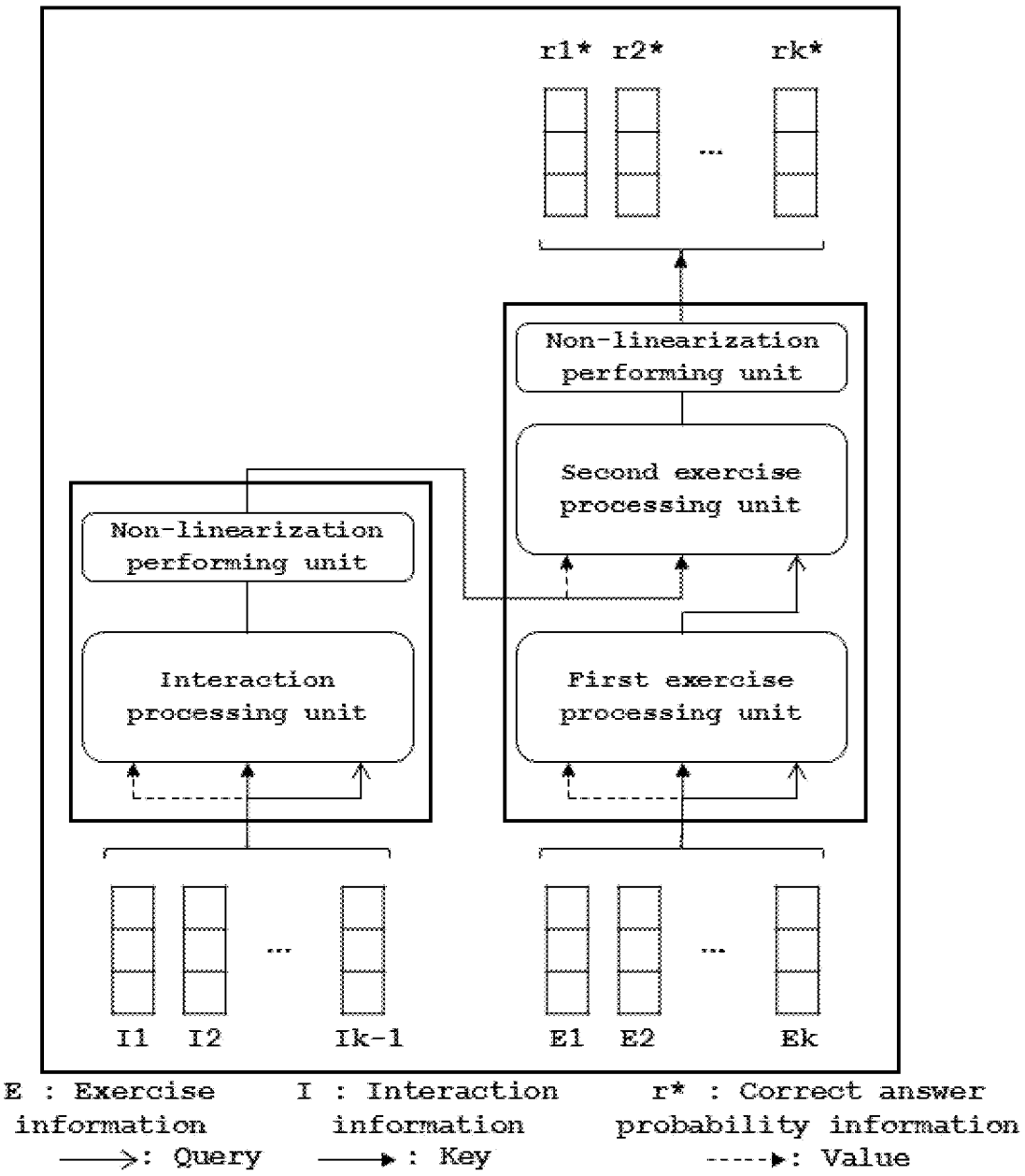

r1* r2* rk*
Non-linearization performing unit
Second exercise processing unit
Non-linearization performing unit
Interaction processing unit
First exercise processing unit
I1 I2 Ik-1
E1 E2 Ek
E : Exercise information
I : Interaction information
r* : Correct answer probability information
: Query
: Key
: Value

FIG 9

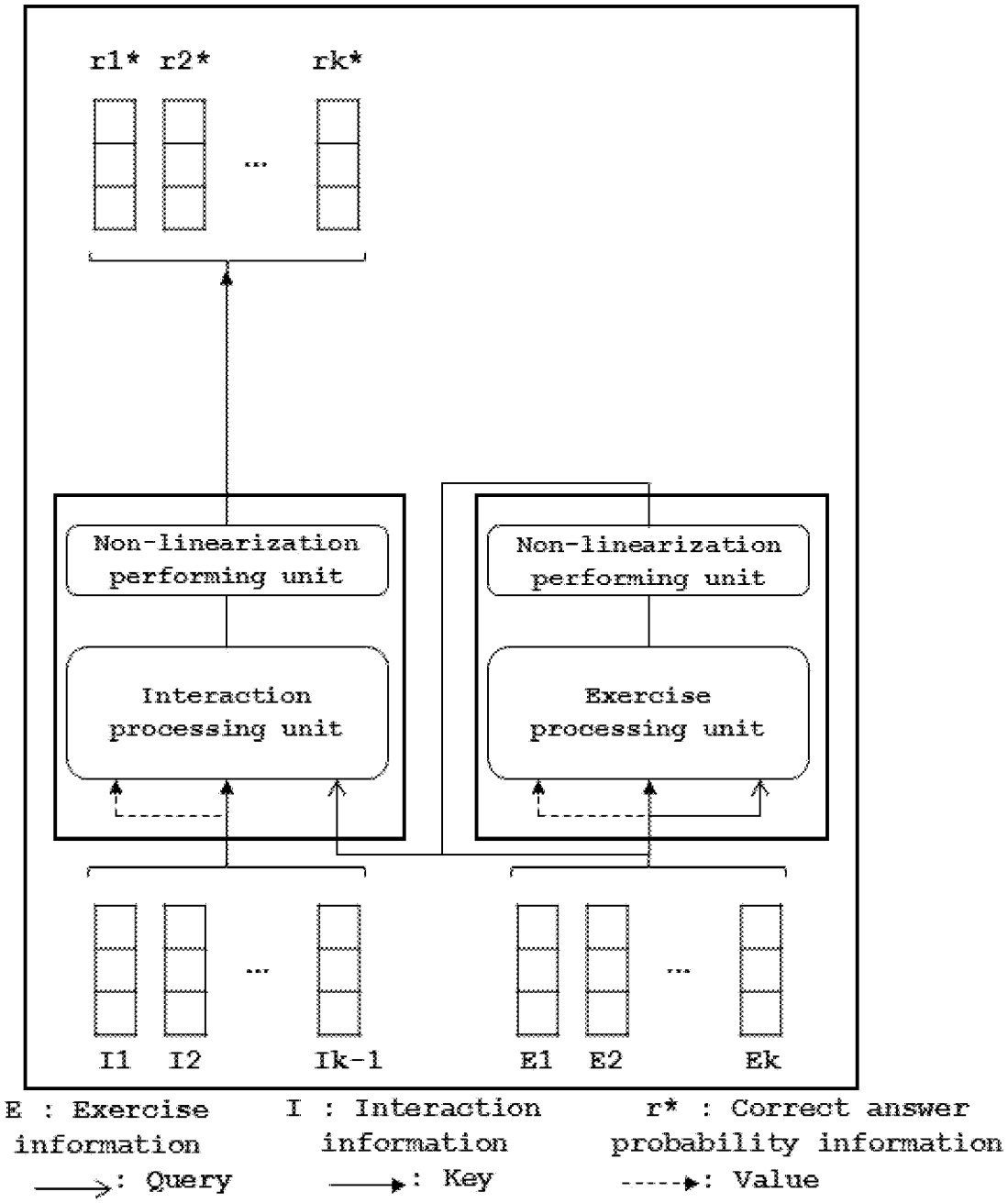

r1* r2* rk*
...
Non-linearization performing unit
Interaction processing unit
Non-linearization performing unit
Exercise processing unit
...
...
I1 I2 Ik-1 E1 E2 Ek
E : Exercise information
I : Interaction information
r* : Correct answer probability information
: Query
: Key
: Value

FIG 10

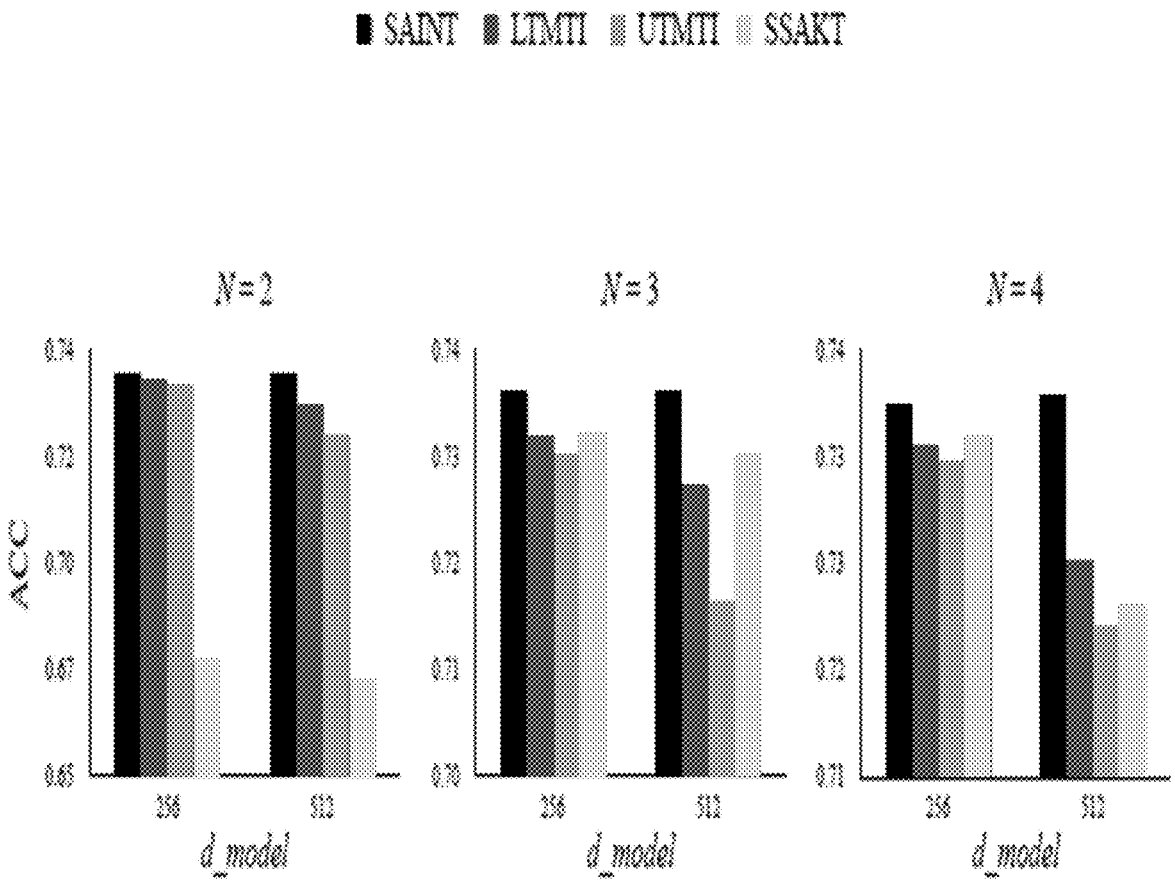

SAINT
LTMTI
UTMTI
SSAKT
N=2
N=3
N=4
ACC
0.74
0.72
0.70
0.67
0.65
0.74
0.73
0.72
0.71
0.70
0.74
0.73
0.72
0.71
256
512
d_model

FIG 11

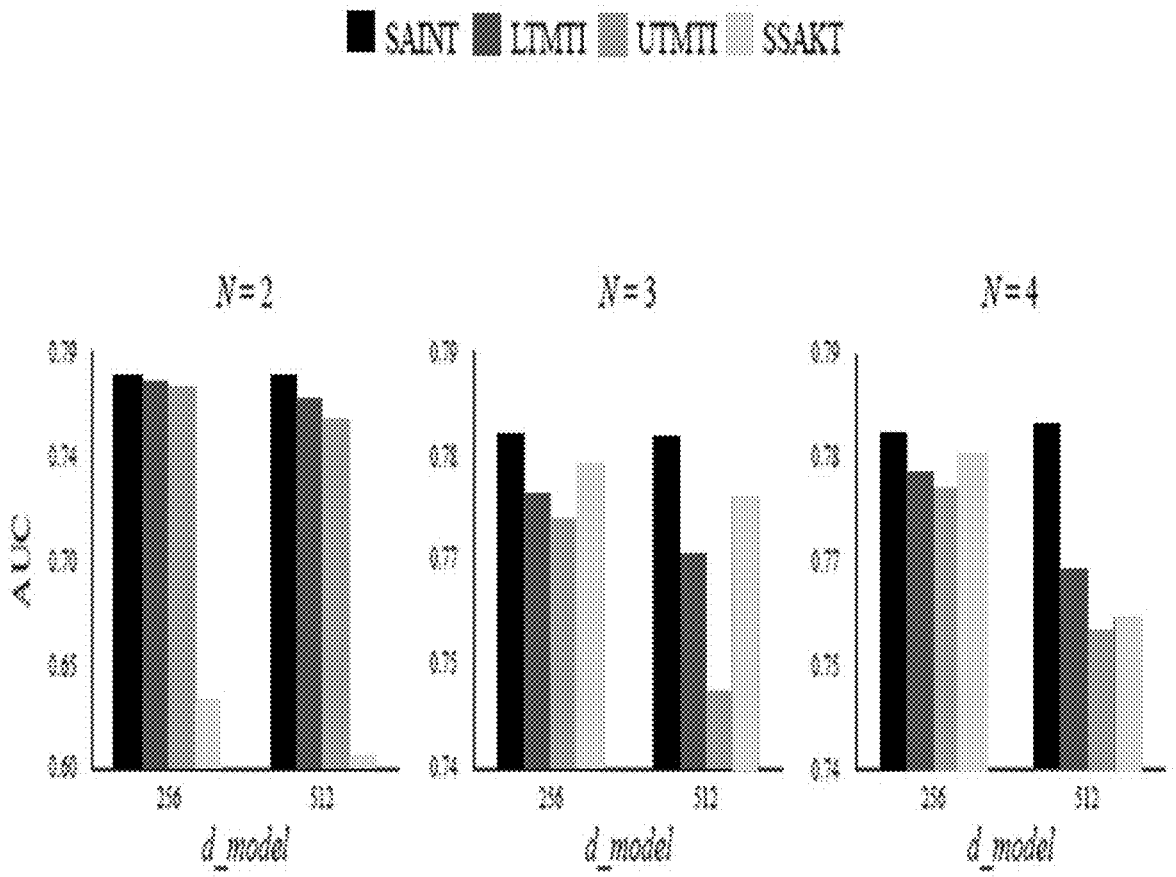

SAINT
LTMTI
UTMTI
SSAKT
N=2
N=3
N=4
AUC
0.80
0.85
0.90
0.74
0.75
0.76
0.77
0.78
0.79
256
512
d_model

FIG 12

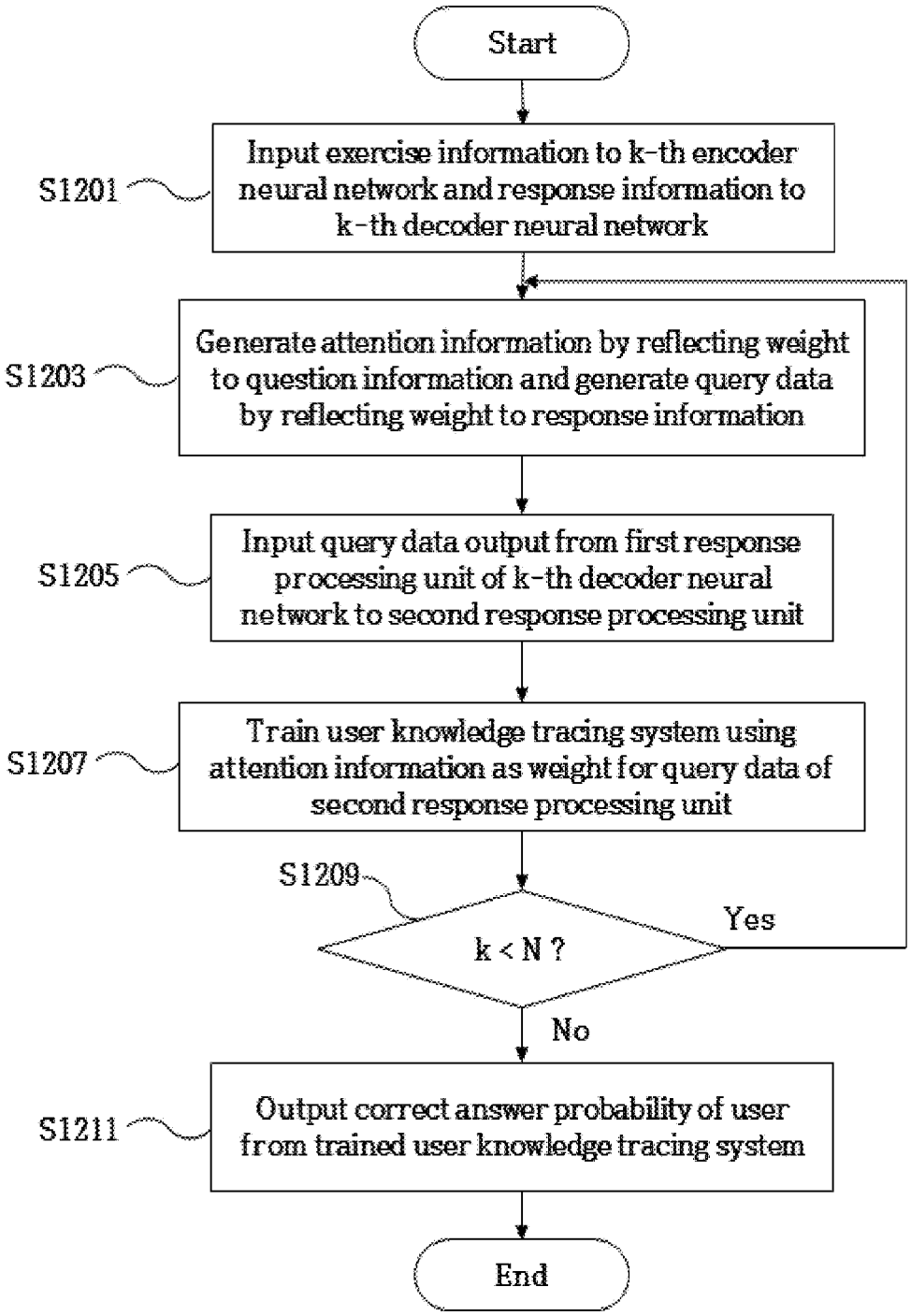
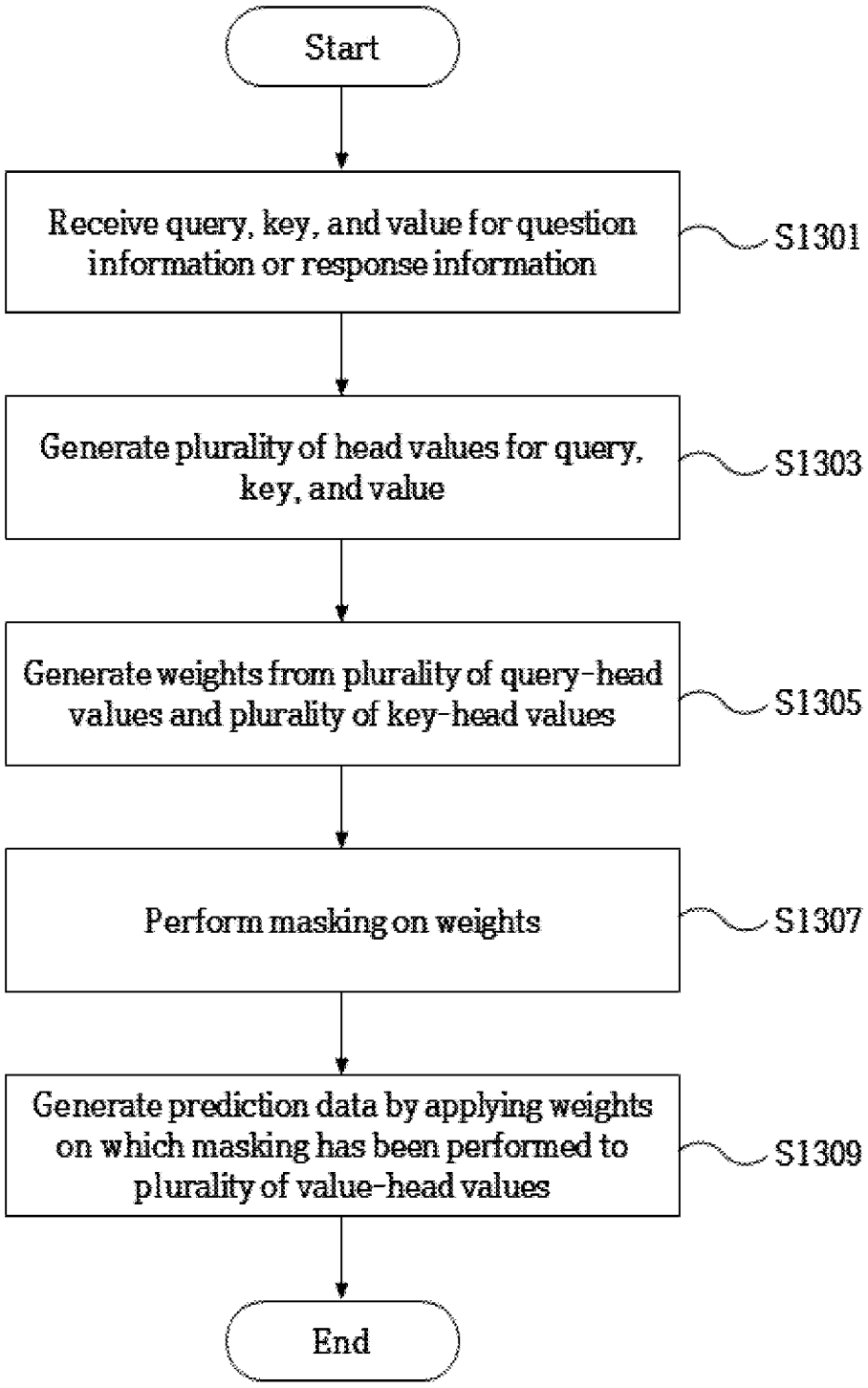

Start
S1201
Input exercise information to k-th encoder neural network and response information to k-th decoder neural network
S1203
Generate attention information by reflecting weight to question information and generate query data by reflecting weight to response information
S1205
Input query data output from first response processing unit of k-th decoder neural network to second response processing unit
S1207
Train user knowledge tracing system using attention information as weight for query data of second response processing unit
S1209
k < N ?
Yes
No
S1211
Output correct answer probability of user from trained user knowledge tracing system
End

FIG 13


Start
Receive query, key, and value for question information or response information
S1301
Generate plurality of head values for query, key, and value
S1303
Generate weights from plurality of query-head values and plurality of key-head values
S1305
Perform masking on weights
S1307
Generate prediction data by applying weights on which masking has been performed to plurality of value-head values
S1309
End

ARTIFICIAL INTELLIGENCE LEARNING-BASED USER KNOWLEDGE TRACING SYSTEM AND OPERATING METHOD THEREOF

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to and the benefit of Korean Patent Application No. 10-2020-0019853, filed Feb. 18, 2020 and Korean Patent Application No. 10-2020-0057446, filed May 14, 2020, the disclosures of which are incorporated herein by reference in their entirety.

BACKGROUND OF THE INVENTION

Field of the Invention

[0002] The present invention relates to a method and system for tracing user knowledge based on artificial intelligence learning. Specifically, the present invention relates to a user knowledge tracing method and system, which predict a correct answer probability of a user by inputting exercise information to an encoder neural network having a transformer architecture and inputting response information to a decoder neural network.

Description of the Related Art

[0003] Recently, the Internet and electronic devices are actively used in various fields, and the educational environment is also rapidly changing. In particular, with the development of various educational media, learners can choose and use a wider range of learning methods. Among them, an education service through the Internet has become a major teaching and learning tool because of advantages of overcoming temporal and spatial constraints and enabling low-cost education.

[0004] This online education service is able to provide more efficient learning content by predicting a user's correct answer probability for a certain exercise, which was not possible in an existing offline education environment, by adopting various artificial intelligence models.

[0005] Conventionally, various artificial neural network models such as RNN, LSTM, bidirectional LSTM, and transformers have been proposed to predict a user's correct answer probability for a given exercise. However, these existing artificial neural network models have a problem in that a layer is too thin to obtain a desired inference result or the input data optimized for user knowledge tracing is not used, so that results with sufficient accuracy cannot be predicted.

[0006] In particular, a transformer model has advantages that a learning speed is very fast and performance is superior to RNN by making an encoder-decoder structure using only attention, not the RNN architecture, so the transformer model begun to attract attention as a user knowledge tracing model, but there is insufficient research to optimize and use the transformer model for learning content.

[0007] Specifically, there is a need for a method for predicting a user's correct answer probability more effectively using a transformer model, with respect to how to configure data input to the encoder and decoder of a transformer model to obtain inference results optimized for learning content, what methods can be used to prevent prediction of the correct answer probability based on an exercise which a user has not yet solved due to the nature of learning content, and the like.

SUMMARY OF THE INVENTION

[0008] The present invention relates to a user knowledge tracing method with more improved accuracy, and an operating method for a user knowledge tracing system including a plurality of encoder neural networks and a plurality of decoder neural networks includes: inputting exercise information to a k-th encoder neural network and inputting response information to a k-th decoder neural network; generating query data, which is information on an exercise for which a user is to predict a correct answer probability by reflecting a weight to the response information and generating attention information to be used as a weight for the query data by reflecting the weight to the exercise information; and training the user knowledge tracing system by using the attention information as the weight for the query data.

[0009] The present invention relates to a user knowledge tracing system having more improved accuracy, and the user knowledge tracing system including a plurality of encoder neural networks and a plurality of decoder neural networks includes: a k-th encoder neural network configured to receive exercise information and generate attention information to be used as a weight for query data by reflecting a weight to the exercise information; and a k-th decoder neural network configured to receive response information, generate the query data, which is information on an exercise for which a user predicts a correct answer probability by reflecting a weight to the response information, and train the user knowledge tracing system using the attention information as a weight for the query data.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] FIG. **1** is a diagram for describing a user knowledge tracing system according to an embodiment of the present invention.

[0011] FIG. **2** is a diagram for describing the operation of the user knowledge tracing system of FIG. **1** in detail.

[0012] FIG. **3** is a diagram for explaining an example of an artificial neural network architecture used in a conventional user knowledge tracing system.

[0013] FIG. **4** is a diagram for explaining another example of an artificial neural network architecture used in a conventional user knowledge tracing system.

[0014] FIG. **5** is a diagram for describing a configuration of each input data.

[0015] FIG. **6** is a diagram for describing key-query masking and upper triangular masking.

[0016] FIG. **7** is a diagram for describing an artificial neural network architecture that outputs a response prediction result using lower triangular masking and interaction information.

[0017] FIG. **8** is a diagram for describing an artificial neural network architecture that outputs a response prediction result using upper triangular masking and interaction information.

[0018] FIG. **9** is a diagram for describing of an artificial neural network architecture that outputs a response prediction result using upper triangular masking, interaction information, and a stacked SAKI.

[0019] FIG. **10** is a graph for comparing ACC performance of the artificial neural network architectures of FIGS. **2** and **7** to **9**.

[0020] FIG. **11** is a graph for comparing AUC performance of the artificial neural network architectures of FIGS. **2** and **7** to **9**.

[0021] FIG. **12** is a flowchart for describing an operation of a user knowledge tracing system according to an embodiment of the present invention.

[0022] FIG. **13** is a flowchart illustrating in detail an operation of an exercise processing unit, a response processing unit, or an interaction processing unit according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0023] Specific structural or step-by-step descriptions of embodiments according to the concept of the present invention disclosed in this specification or application are exemplified only for the purpose of describing the embodiments according to the concept of the present invention, and embodiments according to the concept of the present invention may be implemented in various forms and should not be construed as being limited to the embodiments described in the present specification or application.

[0024] Since the embodiments according to the concept of the present invention can be modified in various ways and have various forms, specific embodiments will be illustrated in the drawings and described in detail in the present specification or application. However, this is not intended to limit the embodiments according to the concept of the present invention to a specific form of disclosure, and it should be understood to include all changes, equivalents, or substitutes included in the spirit and scope of the present invention.

[0025] Terms such as first and/or second may be used to describe various elements, but the elements should not be limited by the terms. The above terms are only for the purpose of distinguishing one component from other components, and a first component may be referred to as a second component, and similarly a second component may also be referred to as a first component, for example, without departing from the scope of claims according to the concept of the present invention.

[0026] It will also be understood that when an element is referred to as being "connected" or "coupled" to another element, it can be directly connected or coupled to the other element or intervening elements may be present. In contrast, when an element is referred to as being "directly connected" or "directly coupled" to another element, there are no intervening elements present. Other expressions describing the relationship between components, such as "between" and "just between" or "adjacent to" and "directly adjacent to" should be interpreted in the same manner.

[0027] Terms used in the disclosure are used to describe specific embodiments and are not intended to limit the scope of the present invention. As used herein, singular forms may include plural forms as well unless the context clearly indicates otherwise. It will be further understood that the terms "comprises," "comprising," "have," "having," "includes," "including" and/or variations thereof, when used in this specification, specify the presence of stated features, numbers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, numbers, steps, operations, elements, components, and/or groups thereof.

[0028] Unless otherwise defined, all terms used herein, including technical or scientific terms, have the same meanings as those generally understood by those skilled in the art to which the present disclosure pertains. Such terms as those defined in a generally used dictionary are to be interpreted as having meanings equal to the contextual meanings in the relevant field of art, and are not to be interpreted as having ideal or excessively formal meanings unless clearly defined in the present application.

[0029] In describing the embodiments, descriptions of technical contents that are well known in the technical field to which the present invention pertains and are not directly related to the present invention will be omitted. This is to more clearly convey the gist of the present invention by omitting unnecessary description.

[0030] Hereinafter, the present invention is described by describing preferred embodiments in detail with reference to the accompanying drawings. Hereinafter, embodiments of the inventive concept will be described in detail with reference to the exemplary drawings.

[0031] FIG. **1** is a diagram for describing a user knowledge tracing system according to an embodiment of the present invention.

[0032] Referring to FIG. **1**, FIG. **1** illustrates a user knowledge tracing system **5** based on a transformer model, and the user knowledge tracing system **5** according to an embodiment may include embedding execution units **10** and **30**, an encoder neural network **20**, and a decoder neural network **40**.

[0033] The transformer model is a model implemented only with attention while complying with the encoder-decoder, which is the architecture of an existing seq2seq. The transformer model does not use an RNN, but maintains an encoder-decoder architecture that receives an input sequence in an encoder and outputs an output sequence in a decoder, like the existing seq2seq, and has characteristics that there may be N units of encoder and decoder.

[0034] The attention mechanism has been proposed to solve the problems of information loss that results from compressing all information into a single vector with a fixed size, and vanishing gradient, which was pointed out as problems of seq2seq based on RNN.

[0035] According to the attention mechanism, every time-step when the decoder predicts an output word, the decoder needs to refer to all of input data from the encoder once again. However, it should be noted that a portion of the input data that is related to data to be predicted at a relevant time point is more attended, rather than referencing all of the input data at the same rate.

[0036] Referring back to FIG. **1**, the user knowledge tracing system **5** trains an artificial neural network based on a large amount of exercise-solving results by the user for an exercise database, and predicts a probability for a correct answer of a specific user for a certain exercise in the exercise database based on the artificial neural network.

[0037] In an educational domain aimed at improving the user's skills, it may be inefficient to provide exercise which the user is sure to correctly answer. It would be efficient to provide an exercise which the user is likely to answer incorrectly, or an exercise which can raise a target test score.

[0038] The user knowledge tracing system **5** according to an embodiment of the present invention may generate a user

model which more accurately reflects user characteristics, and analyze in real time the types of exercises which have high learning efficiency for the user, for example, the exercise types which a user is likely to answer incorrectly, exercises which raise a target test score, or exercises which a user repeatedly answer incorrectly to provide particularly vulnerable exercise types.

[0039] In addition, the user knowledge tracing system **5** may train an artificial neural network based on results of solving a large amount of exercises by the user for an exercise database, and predict a score which a specific user is capable of obtaining on a real test based on the artificial neural network. It is possible to provide user-customized learning according to a predicted score band, and the user can perform more efficient learning.

[0040] Referring to FIG. **1**, the embedding execution units **10** and **30** may perform embedding on input data. In an embodiment of the present invention, the input data may include exercise information, response information, or interaction information.

[0041] The exercise information may be information on exercises having various types and difficulty levels provided to measure a level of knowledge of the user. The response information may be an answer selected by the user in response to the exercise information, or information on whether the user answers a relevant exercise correctly or incorrectly. The interaction information may be information on a set in which the exercise information and the user's response information corresponding thereto are matched.

[0042] In an embodiment, exercise information may be expressed as "E" for Exercise, response information may be expressed as "R" for Response, and interaction information may be expressed as "I" for Interaction. The correct answer probability information may be expressed as 'r*'.

[0043] The embedding execution units **10** and **30** may execute a function of performing embedding in the user knowledge tracing system **5** by expressing input data, for example, an exercise, a response, or a set of an exercise and a response, as vectors. There are many methods to express the input data as vectors for a latent space. For example, it may be one of methods to quantify words and use them in artificial intelligence. Even when the expressions or forms entered by the user are different, the meaning of words, sentences, and texts may be written by calculating association and indicating the association through a numerical value.

[0044] The exercise information expressed as a vector may be embedded in the embedding execution unit **10** and input to the encoder neural network **20**. The response information expressed as a vector may be embedded in the embedding execution unit **30** and input to the decoder neural network **40**.

[0045] According to an embodiment of the present invention, as input data of the transformer model optimized for online learning content, the exercise information is input to the encoder and the response information is input to the decoder, thereby providing the user knowledge tracing system **5** with more improved performance.

[0046] The encoder neural network **20** may generate attention information based on the embedded exercise information. The attention information may be exercise information assigned a weight while passing through a plurality of layers of the encoder neural network **20**. In particular, the attention information may be information generated through self-attention in the encoder neural network. The attention information may be mathematically expressed by a probability, and the sum of all attention information is **1**. The attention information may be input to the decoder neural network **40** and used as a weight for query data of the decoder neural network to be used to train the user knowledge tracing system **5**.

[0047] The artificial neural network may use the attention information to be trained according to an objective function with respect to which part is important. In particular, the self-attention means that attention is performed on oneself, and may be an operation for assigning a weight to a portion to be considered important in specific data itself and reflecting it back to oneself. Since in attention of the existing seq2seq, a correlation has been found with information of different data such as data on the encoder side and data on the decoder side, information according to self-attention is information that cannot be found with the attention architecture of the existing seq2seq.

[0048] In the user knowledge tracing system **5**, the encoder neural network **20** may again refer to all input data E**1**, E**2**, . . . , Ek, R**1**, R**2**, . . . , Rk-**1** at each time-step at which the decoder neural network **40** predicts an output result rk*, and the user knowledge tracing system **5** may focus (attention) on data related to a relevant output result according to the attention information.

[0049] The decoder neural network **40** may generate a response prediction result based on the embedded response information and the attention information. The decoder neural network **40** may perform multi-head attention, in which the above-described self-attention is performed at least once or more, on response information.

[0050] As such, the decoder neural network **40** may generate correct answer probability information by performing multi-head attention on query data generated from the response information, based on the attention information weighted according to a the importance of the exercise information in the encoder neural network **20**.

[0051] According to an embodiment of the present invention, as input data optimized for the user knowledge tracing, the exercise information is input to the encoder and the response information is input to the decoder, thereby providing the user knowledge tracing system **5** with more improved performance.

[0052] In addition, the present invention can provide a user knowledge tracing system **5** having more improved performance by appropriately using upper triangular masking in an encoder neural network and a decoder neural network having a transformer architecture.

[0053] FIG. **2** is a diagram for describing the operation of the user knowledge tracing system of FIG. **1** in detail.

[0054] Referring to FIG. **2**, the user knowledge tracing system **5** may include an encoder neural network **20** and a decoder neural network **40**. Further, the encoder neural network **20** may include an exercise processing unit **21** and a non-linearization performing unit **22**, and the decoder neural network **40** may include a first response processing unit **41**, a second response processing unit **42**, and a non-linearization performing unit **43**.

[0055] Although the embedding performing unit of FIG. **1** is omitted in FIG. **2**, the operation for embedding input data may be understood with reference to the description with reference to FIG. **1** described above.

[0056] The exercise information may be composed of a plurality of exercises E1, E2, . . . , Ek expressed as vectors. The response information may be composed of a user's responses R1, R2, . . . , Rk−1 to the plurality of exercises E1, E2, . . . , Ek expressed as vectors. The correct answer probability information may be composed of the user's correct answer probabilities r1*, r2*, . . . , rk* for the exercises expressed as vectors.

[0057] In an embodiment, the correct answer probability information rk* may be information on a probability that the user will correctly answer an exercise Ek−1 when a user response to the exercise E1 is R1, a user response to the exercise E2 is R2, . . . , and a user response to the exercise Ek−1 is Rk−1.

[0058] The exercise processing unit **21** may receive the exercise information and perform a series of operations related to self-attention. The operation may include dividing the exercise information into queries, keys, and values, generating a plurality of head values for the queries, the keys, and the values, respectively, generating weights from a plurality of query-head values and a plurality of key-head values, performing masking the generated weights, and generating prediction data by applying the masked weights to a plurality of value-head values.

[0059] The prediction data generated by the exercise processing unit **21** may be attention information.

[0060] In particular, the exercise processing unit **21** may perform key-query masking as well as upper triangular masking during the masking operation. The key-query**5** masking and the upper triangular masking will be described in detail in the description with reference to FIG. **6** to be described later.

[0061] The non-linearization performing unit **22** may perform an operation of non-linearizing the prediction data output from the exercise processing unit **21**. An ReLU function can be used for non-linearization.

[0062] Although not shown in the drawing, the encoder neural network **20** may exist at least one or more. Attention information generated by the encoder neural network **20** is input to the encoder neural network **20** again, and a series of operations related to self-attention and non-linearization may be performed repeatedly several times.

[0063] Thereafter, the attention information may be divided into key values and value values and input to the second response processing unit. The attention information may be used as a weight for query data input to the second response processing unit and used to train the user knowledge tracing system.

[0064] The first response processing unit **41** may receive response information and perform a series of operations related to self-attention like the exercise processing unit **21**. The operation may include dividing the exercise information into queries, keys, and values, generating a plurality of head values for the queries, the keys, and the values, generating weights from a plurality of query-head values and a plurality of key-head values, performing masking the generated weights, and generating prediction data by applying the masked weights to a plurality of value-head values.

[0065] The prediction data generated by the first response processing unit **41** may be query data.

[0066] The second response processing unit **42** may receive query data from the first response processing unit and attention information from the encoder neural network **20**, and may output correct answer probability information.

[0067] The attention information may be input to the decoder neural network **40** and used as a weight for query data of the decoder to be used to train the user knowledge tracing system **5**.

[0068] The attention information may be information about a weight assigned to pay attention to a specific area of the query data. Specifically, in the user knowledge tracing system **5**, the encoder neural network **20** may again refer to all input data E1, E2, . . . , Ek, R1, R2, . . . , Rk−1 at each time point at which the decoder neural network **40** predicts an output result rk*, and the user knowledge tracing system **5** may focus (attention) on data related to a relevant output result.

[0069] The second response processing unit **42** may generate rk*, which is the user's correct answer probability information for the exercise information Ek according to the operation.

[0070] Although not shown in the drawing, the decoder neural network **40** may exist at least one or more. The correct answer probability information generated by the decoder neural network **40** may be input to the decoder neural network **40** again, and a series of operations related to self-attention, multi-head attention, and non-linearization may be performed repeatedly several times.

[0071] Like the exercise processing unit **21**, the first response processing unit **41** and the second response processing unit may perform key-query masking as well as upper triangular masking during a masking operation.

[0072] FIG. **3** is a diagram for explaining an example of an artificial neural network architecture used in a conventional user knowledge tracing system.

[0073] In the user knowledge tracing system of FIG. **3**, there is shown an input data processing unit that performs an operation to focus (attention) on a specific portion of input data related to data to be predicted.

[0074] However, in the knowledge tracing system of FIG. **3**, since the layer of the input data processing unit is not deep enough, there is a limitation in which a number of exercises and user responses thereto are not correctly analyzed.

[0075] FIG. **4** is a diagram for explaining another example of an artificial neural network architecture used in a conventional user knowledge tracing system.

[0076] The user knowledge tracing system of FIG. **4** may include an interaction processing unit and a non-linearization performing unit that perform an operation similar to that of the encoder neural network or decoder neural network of FIG. **2**.

[0077] However, in the knowledge tracing system of FIG. **4**, there is a problem in that a limitation of an existing system, in which the interaction information (I) is provided as key and value and the exercise information (E) is provided as query, is not overcame.

[0078] In order to solve such a problem, the user knowledge tracing system according to an embodiment of the present invention can implement an artificial neural network with improved accuracy by predicting a correct answer probability using only exercise information and response information having a smaller amount of data than the interaction information, and sufficiently deeply implementing a layer in which attention is performed.

[0079] FIG. **5** is a diagram for describing configuration of each input data according to an embodiment of the present invention.

[0080] Referring to FIG. **5**, input data may include exercise information (E), interaction information (I), and response information (R). Depending on an artificial neural network model implemented, specific data may be selected and used among three input data.

[0081] Exercise identification information may be a unique value assigned to each exercise. A user or computer may identify what a relevant exercise is through the exercise identification information.

[0082] Exercise category information may be information indicating what type the relevant exercise has. For example, in TOEIC test exercises, an exercise category may be information indicating whether an exercise is a listening part or a reading part.

[0083] Position information may be information indicating where corresponding data is located in the entire data. Since the transformer architecture does not display the sequence of input data unlike the RNN architecture, it is necessary to separately indicate where each data is located in the whole data sequence in order to distinguish the sequence. The position information may be embedded together with the input data, added to the embedded input data, and input to an encoding neural network and a decoding neural network.

[0084] Response accuracy information may be information indicating whether the user's response is a correct answer or an incorrect answer. For example, when the user's response is a correct answer, it may be expressed as a vector representing '1'. Conversely, when the user's response is an incorrect answer, it may be expressed as a vector representing '0'.

[0085] Required time information may be information representing a time required for a user to solve an exercise as a vector. The required time information may be expressed in seconds, minutes, hours, and the like, and it may be determined that the time exceeding a certain time (for example, 300 seconds) has taken as much as the corresponding time (300 seconds).

[0086] Time recording information may be information representing a time point at which a user solves an exercise as a vector. The time recording information may be expressed as time, day, month, year, or the like.

[0087] The exercise information (E) may include the exercise identification information, the exercise category information, and the position information. In other words, the exercise information (E) may include information on what exercise the corresponding exercise is, what type the corresponding exercise has, and where the corresponding exercise is located in the whole exercise data.

[0088] The response information (R) may include the position information and the response accuracy information. That is, the response information (R) may include information on where the user's response is located in the whole response data and whether the user's response is a correct answer or an incorrect answer.

[0089] The interaction information (I) may include the exercise identification information, the exercise category information, the position information, the response accuracy information, the required time information, and the time recording information. The interaction information may additionally include the required time information and the time recording information in addition to all the information of the exercise information (E) and the response information (R).

[0090] The user knowledge tracing system according to an embodiment of the present invention can predict the correct or incorrect answer of the user by using only the exercise information (E) and the response information (R) instead of the interaction information (I) such that the amount of data used is reduced, thus increasing computational performance and having increased memory efficiency. In addition, in terms of accuracy, the exercise information (E) is input to the encoder and the response information (R) is input to the decoder to optimize an online learning environment, thereby predicting the correct answer probability with more improved accuracy.

[0091] FIG. **6** is a diagram for describing key-query masking and upper triangular masking.

[0092] Although FIG. **6** shows that the upper triangular masking is performed after the key-query masking, the upper triangular masking and the key-query masking both may be performed simultaneously, or the upper triangular masking may be performed first.

[0093] The key-query masking is optional, and may be an operation in which attention is not performed by imposing a penalty on a value without a value (zero padding). A value of prediction data on which key-query masking has been performed may be expressed as 0, and the remaining portion may be expressed as 1.

[0094] Although, in the key-query masking of FIG. **6**, the last values of the Query and the Key are masked for convenience of description, this may be variously changed according to embodiments.

[0095] Upper triangular masking may be an operation for preventing attention from being performed on information corresponding to a future position for prediction of a next exercise. For example, it may be an operation for performing masking to prevent a prediction value from being calculated from an exercise that the user has not yet solved. Like the key-query masking, a value of prediction data on which upper triangular masking has been performed may be expressed as 0, and the remaining portion may be expressed as 1.

[0096] Values of the masked prediction data may be controlled to have a probability close to zero when a random large negative value is reflected and expressed probabilistically through a soft-max function.

[0097] In the conventional transformer architecture, the key-query masking is performed in the encoder neural network, and upper triangular masking is performed in addition to the key-query masking in the decoder neural network. In an embodiment of the present invention, by performing upper triangular masking in both the encoder neural network and the decoder neural network, it can be controlled such that correct answer probability information depends only on exercise information (E**1**, E**2**, . . . , Ek) previously provided to the user and response information (R**1**, R**2**, . . . , Rk−1) previously submitted by the user.

[0098] FIG. **7** is a diagram for describing an artificial neural network architecture that outputs a response prediction result using lower triangular masking and interaction information.

[0099] Referring to FIG. **7**, interaction information may be input to an encoder neural network as input data, and exercise information may be input to a decoder neural network as input data. In addition, the decoder neural network may only perform multi-head attention in which self-attention is omitted.

[0100] Furthermore, the interaction processing unit and the exercise processing unit of FIG. **7** may perform attention in which an upper triangle of prediction data is not masked, but a lower triangle is masked.

[0101] FIG. **8** is a diagram for describing an artificial neural network architecture (UTMTI) that outputs a response prediction result using upper triangular masking and interaction information.

[0102] Referring to FIG. **8**, interaction information may be input to an encoder neural network as input data, and exercise information may be input to a decoder neural network as input data.

[0103] FIG. **9** is a diagram for describing an artificial neural network architecture (SSAKT) that outputs a response prediction result using upper triangular masking, interaction information, and stacked SAKT.

[0104] Referring to FIG. **9**, interaction information may be input to an encoder neural network as input data, and exercise information may be input to a decoder neural network as input data. In addition, the interaction processing unit may perform multi-head attention instead of self-attention, and the response processing unit may perform self-attention.

[0105] FIG. **10** is a graph for comparing ACC performance of the artificial neural network architectures of FIGS. **2** and **7** to **9**.

[0106] ACC can be an indicator of sensitivity. ACC may indicate the proportion of response information that is a correct answer among all response information with an incorrect answer. N may denote the number of stacked encoders and decoders. d_model may denote the output order of all lower layers of the model. The user knowledge tracing system of FIG. **2** is indicated by SAINT.

[0107] As a result of statistics, it can be seen that, on average, the ACC of SAINT is about 1.8% higher than that of other models.

[0108] FIG. **11** is a graph for comparing AUC performance of the artificial neural network architectures of FIGS. **2** and **7** to **9**.

[0109] AUC may represent the ratio of the correct prediction to the overall prediction. N may denote the number of stacked encoders and decoders. d_model may denote the output order of all lower layers of the model. The user knowledge tracing system of FIG. **2** is indicated by SAINT.

[0110] As a result of statistics, it can be seen that, on average, the AUC of SAINT is about 1.07% higher than that of other models.

[0111] FIG. **12** is a flowchart for describing an operation of a user knowledge tracing system according to an embodiment of the present invention.

[0112] Referring to FIG. **12**, in step **S1201**, the user knowledge tracing system may input exercise information to a k-th encoder neural network and response information to a k-th decoder neural network, respectively.

[0113] The exercise information (E) may include the exercise identification information, the exercise category information, and the position information. In other words, the exercise information (E) may include information on what exercise the corresponding exercise is, what type the exercise has, and where the exercise is located in the whole exercise data.

[0114] The response information (R) may include the position information and the response accuracy information. That is, the response information (R) may include information on where the user's response is located in the whole response data and whether the user's response is a correct answer or an incorrect answer.

[0115] In step **S1203**, the user knowledge tracing system may generate attention information by reflecting a weight to the exercise information, and generate query data by reflecting a weight to the response information.

[0116] Specifically, the weight of the exercise information may be reflected to the exercise information itself through self-attention. The self-attention may be an operation for assigning a weight to a portion to be considered important in specific data itself and reflecting it back to oneself.

[0117] Weights may be reflected on the response information by performing not only self-attention but also multi-head attention based on the attention information.

[0118] In step **S1205**, query data output from the first response processing unit of the k-th decoder neural network may be input to the second response processing unit. The query data may be prediction data output from the first response processing unit.

[0119] In step **S1207**, the user knowledge tracing system may train the user knowledge tracing system by using the attention information as a weight for query data of the second response processing unit.

[0120] In step **S1209**, the user knowledge tracing system may compare k and N, and when k is greater than or equal to N, performs step **S1211**, and when k is less than N, returns to step **S1203** and performs steps **S1203** to **S1207** again.

[0121] Since the encoder neural network and the decoder neural network can be stacked as many as N, the above process can be repeatedly performed for all of the stacked encoder neural networks and the stacked decoder neural networks until the operation is finished.

[0122] In step **S1211**, the user knowledge tracing system may output the user's correct answer probability information from the trained user knowledge tracing system.

[0123] This is an inference process, and it is possible to process input data according to a weight determined in a learning process, and output correct answer probability information indicating a correct answer probability of an exercise solved by a user.

[0124] FIG. **13** is a flowchart for describing, in detail, an operation of an exercise processing unit, a response processing unit, or an interaction processing unit according to an embodiment of the present invention.

[0125] Referring to FIG. **13**, in step **S1301**, each component may receive query, key, and value for exercise information or response information.

[0126] The value of each of the query, the key, and the value may be a value expressed as a vector, and may be a value classified according to a role used.

[0127] In step **S1303**, each component may generate a plurality of head values for each of the query, the key, and the value.

[0128] In step **S1305**, each component may generate weights from a plurality of query-head values and a plurality of key-head values, and in step **S1307**, masking operation including key-query masking and upper triangular masking may be performed.

[0129] The key-query masking is optional, and may be an operation in which attention is not performed by imposing a penalty on a value without a value (zero padding). A value of prediction data on which upper triangular masking has

been performed may be expressed as 0, and the remaining portion may be expressed as 1.

[0130] Upper triangular masking may be an operation for preventing attention from being performed on information corresponding to a future position for prediction of a next exercise. For example, it may be an operation for performing masking to prevent a predicted value from being calculated from an exercise that the user has not yet solved. Like the key-query masking, a value of prediction data on which upper triangular masking has been performed may be expressed as 0, and the remaining portion may be expressed as 1.

[0131] Thereafter, in step S**1309**, prediction data may be generated by applying the masked weight to a plurality of value-head values.

[0132] The prediction data generated by the exercise processing unit may be attention information, the prediction data generated by the first response processing unit may be query data, and the prediction data generated by the second response processing unit may be correct answer probability information.

[0133] The user knowledge tracing system according to the present invention can have improved performance by using an optimized input data format and appropriately using upper triangular masking for an encoder neural network and a decoder neural network having a transformer architecture.

[0134] The present invention is to solve the above-described problem, and provides a user knowledge tracing system having improved performance by using an input data format optimized for user knowledge tracing.

[0135] In addition, the present invention provides a user knowledge tracing system with improved performance by appropriately using upper triangular masking in an encoder neural network and a decoder neural network having a transformer architecture.

[0136] The embodiments of the present invention disclosed in the present specification and drawings are provided only to provide specific examples to easily describe the technical contents of the present invention and to aid understanding of the present invention, and are not intended to limit the scope of the present invention. It is obvious to those of ordinary skill in the art that other modifications based on the technical idea of the invention can be implemented in addition to the embodiments disclosed therein.

What is claimed is:

1. An operating method for a user knowledge tracing system including a plurality of encoder neural networks and a plurality of decoder neural networks, the operating method comprising:

inputting exercise information to a k-th encoder neural network and inputting response information to a k-th decoder neural network;

generating query data, which is information on an exercise for which a user is to predict a correct answer probability by reflecting a weight to the response information and generating attention information to be used as a weight for the query data by reflecting the weight to the exercise information; and

training the user knowledge tracing system by using the attention information as the weight for the query data.

2. The user knowledge tracing system of claim **1**, further comprising:

in a case where a number of the plurality of encoder neural networks and the plurality of decoder neural networks which are stacked is N, when k is less than N, repeatedly performing the step of generating the attention information and the step of training the user knowledge tracing system.

3. The user knowledge tracing system of claim **2**, further comprising:

when k is equal to or greater than N, completing training of the user knowledge tracing system and outputting correct answer probability information, which is a probability that a user correctly answers an exercise, from the trained user knowledge tracing system.

4. The user knowledge tracing system of claim **1**, wherein the generating of the attention information includes

optionally performing key-query masking, which is an operation for preventing attention from being performed on a value without a value (zero padding); and

performing upper triangular masking, which is an operation for preventing attention from being performed on information corresponding to a future position for prediction of a next exercise.

5. The user knowledge tracing system of claim **1**,

wherein the exercise information is composed of a plurality of exercises expressed as vectors, wherein the response information is composed of responses of the user to the plurality of exercises expressed as vectors.

6. A user knowledge tracing system including a plurality of encoder neural networks and a plurality of decoder neural networks, comprising:

a k-th encoder neural network configured to receive exercise information and generate attention information to be used as a weight for query data by reflecting a weight to the exercise information; and

a k-th decoder neural network configured to receive response information, generate the query data, which is information on an exercise for which a user predicts a correct answer probability, by reflecting a weight to the response information, and train the user knowledge tracing system using the attention information as a weight for the query data.

* * * * *