

METHOD, APPARATUS AND DEVICE FOR PREDICTING PRONUNCIATION OF POLYPHONE, AND STORAGE MEDIUM

The present application claims priority to Chinese Patent Application No. 202011432585.6, filed with the China National Intellectual Property Administration on December 10, 2020, and entitled "METHOD, APPARATUS AND DEVICE FOR PREDICTING PRONUNCIATION OF POLYPHONE, AND STORAGE MEDIUM", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

The present application relates to the field of intelligent decision making of artificial intelligence, and in particular, to a method, apparatus and device for predicting a pronunciation of a polyphone, and a storage medium.

BACKGROUND

Grapheme-to-phoneme conversion is an important part of a text-to-speech system. But unlike other languages, it is very common for a character in Chinese to have different pronunciations under different circumstances, and many Chinese characters even have more than 3 pronunciations. Therefore, quality of a system for polyphone pronunciation labeling greatly affects quality of a Chinese speech synthesis system. If a pronunciation is incorrectly labeled, an obvious error in synthesized speech is caused. At present, in methods for predicting a pronunciation of a polyphone, a pronunciation of a polyphone is predicted usually by using labeled data and randomly initializing a set of vectors of the labeled data.

However, the inventor realized that, due to the random initialization of a set of vectors, a phrase that has not been labeled during model training cannot be recognized in prediction of a pronunciation of a polyphone, i.e., the out-of-vocabulary problem occurs, resulting in relatively low accuracy of predicting a pronunciation of a polyphone.

SUMMARY

The present application provides a method, apparatus and device for predicting a pronunciation of a polyphone, and a storage medium to improve accuracy of predicting a pronunciation of a polyphone.

According to a first aspect of the present application, a method for predicting a pronunciation of a polyphone is provided, including:

obtaining a labeled to-be-processed Chinese sentence, and obtaining a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone;

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability.

According to a second aspect of the present application, an apparatus for predicting a pronunciation of a polyphone is provided, including:

an obtaining module, configured to obtain a labeled to-be-processed Chinese sentence, and obtain a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone;

a conversion module, configured to perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and convert the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

a splicing module, configured to perform splicing processing on the polyphone

representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

a determining module, configured to calculate a target pinyin probability of the target vector through a preset linear layer, and determine a target pronunciation of the target polyphone based on the target pinyin probability.

According to a third aspect of the present application, a device for predicting a pronunciation of a polyphone is provided, including: a memory and at least one processor, where the memory stores instructions; and the at least one processor invokes the instructions in the memory to enable the device for predicting a pronunciation of a polyphone to perform the method for predicting a pronunciation of a polyphone including the following steps:

obtaining a labeled to-be-processed Chinese sentence, and obtaining a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone;

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability.

According to a fourth aspect of the present application, a computer-readable storage medium is provided, where the computer-readable storage medium store instructions, and when the instructions are run on a computer, the computer is enabled to perform the method for predicting a pronunciation of a polyphone including the following steps:

obtaining a labeled to-be-processed Chinese sentence, and obtaining a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone;

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability.

In the technical solutions according to the present application, a labeled to-be-processed Chinese sentence is obtained, and a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence are obtained, where the to-be-processed Chinese sentence includes a target polyphone; phrase segmentation processing is performed on the to-be-processed Chinese sentence to obtain a target segmented phrase, and the character representation vector set is converted into a phrase-level feature representation vector based on the target segmented phrase; splicing processing is performed on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and a target pinyin probability of the target vector is calculated through a preset linear layer, and a target pronunciation of the target polyphone is determined based on the target pinyin probability. In the embodiments of the present application, a character representation vector set is converted into a phrase-level feature representation vector based on a target segmented phrase, and character features are converted to phrase-level features, so as to avoid the out-of-vocabulary problem, thereby effectively improving accuracy of predicting a pronunciation of a polyphone. Splicing processing based on an attention mechanism is performed on a polyphone representation vector and the phrase-level feature representation vector. A target pinyin probability of a target vector is calculated through a preset linear layer. A target pronunciation of a target polyphone is determined based on the target pinyin probability. A pronunciation of the target polyphone is predicted by combining the target segmented phrase and the attention mechanism. Such practice reduces impact of the labeling error problem existing in

segmented phrases without any rule or artificial feature design, can accurately capture text semantic information of a to-be-processed Chinese sentence, and improves the accuracy of predicting a pronunciation of a polyphone.

BRIEF DESCRIPTION OF DRAWINGS

FIG. 1 is a schematic diagram of a method for predicting a pronunciation of a polyphone according to an embodiment of the present application;

FIG. 2 is a schematic diagram of a method for predicting a pronunciation of a polyphone according to another embodiment of the present application;

FIG. 3 is a schematic diagram of an apparatus for predicting a pronunciation of a polyphone according to an embodiment of the present application;

FIG. 4 is a schematic diagram of an apparatus for predicting a pronunciation of a polyphone according to another embodiment of the present application; and

FIG. 5 is a schematic diagram of a device for predicting a pronunciation of a polyphone according to an embodiment of the present application.

DESCRIPTION OF EMBODIMENTS

Embodiments of the present application provide a method, apparatus and device for predicting a pronunciation of a polyphone, and a storage medium to improve accuracy of predicting a pronunciation of a polyphone.

Terms "first", "second", "third", "fourth", etc. (if any) in the specification, claims, and accompanying drawings of the present application are used to distinguish between similar objects without having to describe a specific order or sequence. It should be understood that data used in this way may be interchanged under appropriate circumstances, so that the embodiments described herein can be implemented in an order other than that illustrated or described herein. In addition, the term "including" or "having" and any variants thereof are intended to cover non-exclusive inclusions. For example, a process, method, system, product, or device that includes a series of steps or units is not necessarily limited to those steps or units clearly listed, and may include other steps or units that are not clearly listed or are inherent to the process,

method, product, or device.

For ease of understanding, a specific process of an embodiment of the present application is described below. Referring to FIG. 1, a method for predicting a pronunciation of a polyphone in an embodiment of the present application includes the following steps.

101: Obtain a labeled to-be-processed Chinese sentence, and obtain a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone.

It is understandable that an execution body of the present application may be an apparatus for predicting a pronunciation of a polyphone or a terminal or a server, which is not specifically limited herein. In this embodiment of the present application, that the server is the execution body is used as an example for description.

A server receives an initial Chinese sentence sent by a preset graphical interface, performs data cleansing on the initial Chinese sentence to obtain a candidate Chinese sentence, and obtains a pre-created polyphone label. The polyphone label may be a note created based on a polyphone in at least one of a general dictionary, a business field dictionary and user profile tags to improve universality and accuracy of polyphone labeling based on multiple fields, and improve labeling accuracy of the polyphone through interest and hobbies based on the user profile tags. The polyphone label includes a polyphone and a pronunciation of the polyphone based on semantic information. The server recognizes a business field and user information of the candidate Chinese sentence, invokes a corresponding polyphone label based on the business field and the user information, recognizes a target polyphone in the candidate Chinese sentence through the polyphone label, and labels the target polyphone, so to obtain the labeled to-be-processed Chinese sentence.

After obtaining the labeled to-be-processed Chinese sentence, the server invokes a pre-trained character vector and a preset character vector conversion algorithm to perform vector conversion on characters in the to-be-processed Chinese sentence to obtain a character representation vector set, and extracts a representation vector corresponding to the target

polyphone in the character representation vector set based on the labeled target polyphone to obtain a polyphone representation vector; or the server extracts the target polyphone in the labeled to-be-processed Chinese sentence, and invokes the pre-trained character vector and the preset character vector conversion algorithm to separately perform vector conversion on the characters in the to-be-processed Chinese sentence and the target polyphone, so as to obtain a character representation vector set and a polyphone representation vector. There is one or more target polyphones.

102: Perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and convert the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase.

The server invokes a preset Jieba phrase segmentation tool or a Chinese language processing package HanLP phrase segmentation tool or another phrase segmentation tool to perform phrase segmentation processing on the to-be-processed Chinese sentence based on an original sentence sequence to obtain an initial segmented phrase; or the server invokes a preset dictionary-based Chinese phrase segmentation algorithm or a statistics-based Chinese phrase segmentation algorithm to perform phrase segmentation processing on the to-be-processed Chinese sentence based on an original sentence sequence to obtain an initial segmented phrase, and performs splicing on the initial segmented phrase based on a preset phrase splicing rule to obtain the target segmented phrase. There is one or more initial segmented phrases and one or more target segmented phrases. The server classifies a character representation vector in the character representation vector set based on the target segmented phrase to obtain a character representation vector group corresponding to each target segmented phrase, and splices the character representation vector group corresponding to each target segmented phrase to obtain the phrase-level feature representation vector. There is one or more phrase-level feature representation vectors, and one target segmented phrase corresponds to one phrase-level feature representation vector.

103: Perform splicing processing on the polyphone representation vector and the

phrase-level feature representation vector based on an attention mechanism to obtain a target vector.

The server may, through the preset attention mechanism, calculate a polyphone attention value of the polyphone representation vector, multiply the polyphone attention value by the polyphone representation vector to obtain a polyphone vector matrix, calculate a phrase attention value of the phrase-level feature representation vector that is based on the polyphone representation vector to obtain a phrase vector matrix, and perform matrix addition or matrix multiplication on the polyphone vector matrix and the phrase vector matrix to obtain the target vector; or the server may, through the preset attention mechanism, calculate a first attention value of the polyphone representation vector relative to the phrase-level feature representation vector, calculate a second attention value of the phrase-level feature representation vector relative to the polyphone representation vector, multiply the first attention value by the phrase-level feature representation vector to obtain a first vector, multiply the second attention value by the polyphone representation vector to obtain a second vector, and perform matrix addition or matrix multiplication on the first vector and the second vector to obtain the target vector.

104: Calculate a target pinyin probability of the target vector through a preset linear layer, and determine a target pronunciation of the target polyphone based on the target pinyin probability.

There may be multiple preset linear layers, and each layer corresponds to one classifier, i.e., the linear layers include a plurality of classifiers. The server performs pinyin classification and probability value calculation on the target vector through the plurality of classifiers to obtain a plurality of initial pinyin probabilities corresponding to each classifier, and performs weighted summation on the initial pinyin probabilities corresponding to the plurality of classifiers to obtain a target pinyin probability of the target vector. There is one or more initial pinyin probabilities. The server compares the initial pinyin probabilities with a preset threshold and compares values of the initial pinyin probabilities for analysis to obtain the target pinyin probability, and determines a pinyin corresponding to the target pinyin probability as a target pronunciation of the

target polyphone. For example, the plurality of classifiers are classifier 1, classifier 2 and classifier 3. Classifier 1 performs pinyin classification and probability value calculation on the target vector to obtain probability A1 based on pinyin 1 and probability A2 based on pinyin 2; classifier 2 performs pinyin classification and probability value calculation on the target vector to obtain probability B1 based on pinyin 1 and probability B2 based on pinyin 2; and classifier 3 performs pinyin classification and probability value calculation on the target vector to obtain probability C1 based on pinyin 1 and probability C2 based on pinyin 2. Weighted summation is performed on A1, B1 and C1 to obtain initial pinyin probability 1 of the target vector based on pinyin 1; and weighted summation is performed on A2, B2 and C2 to obtain initial pinyin probability 2 of the target vector based on pinyin 2. If one of initial pinyin probability 1 and initial pinyin probability 2 is greater than the preset threshold, the initial pinyin probability greater than the preset threshold is determined as the target pinyin probability. If initial pinyin probability 1 and initial pinyin probability 2 are both greater than the preset threshold, the larger one is determined as the target pinyin probability. If initial pinyin probability 1 and initial pinyin probability 2 are both less than or equal to the preset threshold, the initial pinyin probability is recalculated. After obtaining the target pinyin probability, the server determines a pinyin corresponding to the target pinyin probability as the target pronunciation of the target polyphone.

In another embodiment, the server performs matching against initial historical polyphone information stored in a preset database based on the to-be-processed Chinese sentence and the target polyphone to obtain corresponding target historical polyphone information, where the target historical polyphone information includes a target historical Chinese sentence, a historical polyphone in the target historical Chinese sentence and a pronunciation of the historical polyphone; calculates a similarity between the target pronunciation of the target polyphone and the pronunciation of the historical polyphone; calculates a difference between the similarity and 1 to obtain a target value, determines whether the target value is less than a preset similarity value, if yes, determines the target pronunciation of the target polyphone as a final target pronunciation, and if not, determines the pronunciation of the historical polyphone as the target pronunciation of

the target polyphone.

In the embodiment of the present application, a character representation vector set is converted into a phrase-level feature representation vector based on a target segmented phrase, and character features are converted to phrase-level features, so as to avoid the out-of-vocabulary problem, thereby effectively improving accuracy of predicting a pronunciation of a polyphone. Splicing processing based on an attention mechanism is performed on a polyphone representation vector and the phrase-level feature representation vector. A target pinyin probability of a target vector is calculated through a preset linear layer. A target pronunciation of a target polyphone is determined based on the target pinyin probability. A pronunciation of the target polyphone is predicted by combining the target segmented phrase and the attention mechanism. Such practice reduces impact of the labeling error problem existing in segmented phrases without any rule or artificial feature design, can accurately capture text semantic information of a to-be-processed Chinese sentence, and improves the accuracy of predicting a pronunciation of a polyphone.

Referring to FIG. 2, a method for predicting a pronunciation of a polyphone according to another embodiment of the present application includes:

201: Obtain a labeled to-be-processed Chinese sentence, and obtain a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone.

Specifically, the server obtains an initial Chinese sentence, a target polyphone in the initial Chinese sentence and polyphone position information corresponding to the target polyphone; labels the target polyphone in the initial Chinese sentence based on the polyphone position information to obtain the to-be-processed Chinese sentence; and sequentially performs character vector encoding and polyphone vector extraction on the to-be-processed Chinese sentence to obtain the character representation vector set and the polyphone representation vector.

The server receives the initial Chinese sentence sent by a preset graphical interface, invokes a pre-created polyphone dictionary to perform polyphone recognition on the initial Chinese sentence to obtain the target polyphone, extracts position information (i.e., polyphone position

information) of the target polyphone in the initial Chinese sentence, and labels the target polyphone corresponding to the polyphone position information in the initial Chinese sentence. Labeled content includes the target polyphone and the polyphone position information of the target polyphone, and may further include a pronunciation of the target polyphone based on a Chinese sentence corresponding to the initial Chinese sentence. The server may perform matching against the Chinese sentence corresponding to the initial Chinese sentence by calculating a weighted sum value of a semantic similarity, an emotional similarity and a sentence pattern similarity to obtain the to-be-processed Chinese sentence.

The server invokes a preset supervised neural network encoder and/or unsupervised pre-trained network encoder to perform character vector encoding on the to-be-processed Chinese sentence to obtain the character representation vector set, and extracts the polyphone representation vector corresponding to the target polyphone from the character representation vector set.

Specifically, the server encodes each character in the to-be-processed Chinese sentence by using a preset deep neural network encoder to obtain the character representation vector set, where one character representation vector corresponds to one character; and extracts a representation vector corresponding to the target polyphone from the character representation vector set based on the polyphone position information, so as to obtain the polyphone representation vector.

The server invokes the deep neural network encoder in the preset supervised neural network encoder. The deep neural network encoder may include, but is not limited to, at least one of a long short-term memory (LSTM) model and a transformer-based bidirectional encoder representations from transformers (BERT) model. Through the deep neural network encoder, encoding based on contextual semantic information is performed on each character in the to-be-processed Chinese sentence based on the sequence of the characters in the to-be-processed Chinese sentence, so as to obtain a representation vector of each character, i.e., the character representation vector set, and a representation vector corresponding to the polyphone position

information in the character representation vector set is extracted to obtain the polyphone representation vector. For example, the to-be-processed Chinese sentence is "所有商品都打折出售", the polyphone position information indicates the seventh character in the to-be-processed Chinese sentence, and a seventh-character representation vector is extracted from the character representation vector set to obtain a polyphone representation vector corresponding to the target polyphone.

202: Perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and convert the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase.

Specifically, the server performs phrase segmentation processing on the to-be-processed Chinese sentence to obtain the target segmented phrase; divides the character representation vector set based on the target segmented phrase to obtain a representation vector set of each phrase; and performs hybrid pooling on the representation vector set of each phrase through a preset hybrid pooling layer to obtain the phrase-level feature representation vector.

The server invokes a preset Chinese phrase segmentation algorithm to perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain an initial segmented phrase, performs part-of-speech detection and phrase detection on the initial segmented phrase and determines an initial segmented phrase that has passed the detection as the target segmented phrase. The Chinese phrase segmentation algorithm combines an N-gram model and a bi-direction matching method (BM) model, i.e., an output of the N-gram model may be an input of the BM model, or an output of the BM model may be an input of the N-gram model, or, the N-gram model and the BM model are connected in parallel.

The server divides the character representation vector set based on the target segmented phrase to obtain a representation vector set of each phrase. For example, the to-be-processed Chinese sentence is "所有商品都打折出售", and corresponding target segmented phrases are "所有", "商品", "都", "打折" and "出售". Taking "打折" as an example, the representation vector set of the phrase "打折" includes a representation vector of "打" and a representation vector of "

折", and the remaining phrases are described in the same way.

A preset hybrid pooling layer is used to indicate a pooling layer that combines max pooling and average pooling. The server invokes the preset hybrid pooling layer to perform hybrid pooling on the representation vector group of each phrase to obtain the phrase-level feature representation vector. For example, the representation vector of "打" and the representation vector of "折" in the representation vector group of the phrase are fused to obtain the phrase-level feature representation vector of "打折". The server may use a max pooling convolution kernel or a max pooling layer in the hybrid pooling layer to perform max pooling on the representation vector group of each phrase to obtain a first phrase representation vector group, and use an average pooling convolution kernel or an average pooling layer in the hybrid pooling layer to perform average pooling on the first phrase representation vector group to obtain a phrase-level feature representation vector; or the server uses the max pooling convolution kernel or the max pooling layer in the hybrid pooling layer to perform max pooling on the representation vector group of each phrase to obtain a first phrase representation vector group, uses the average pooling convolution kernel or the average pooling layer in the hybrid pooling layer to perform average pooling on the representation vector group of each phrase to obtain a second phrase representation vector group, fuses the first phrase representation vector group and the second phrase representation vector group to obtain a phrase-level feature representation vector; or the server pre-creates a hybrid pooling layer that fuses the max pooling convolution kernel and the average pooling convolution kernel, performs pooling convolution processing on the representation vector group of each phrase to obtain a phrase-level feature representation vector. There is one or more phrase-level feature representation vectors, and one target segmented phrase corresponds to one phrase-level feature representation vector.

203: Perform attention calculation on the polyphone representation vector and the phrase-level feature representation vector through a preset feed-forward attention mechanism to obtain an attention vector.

The server uses a preset feed-forward attention mechanism to calculate an attention value

between the polyphone representation vector and the phrase-level feature representation vector, and performs weighed summation on the polyphone representation vector and the phrase-level feature representation vector through the attention value to obtain an attention vector; or the server uses the preset feed-forward attention mechanism to calculate an attention value of the polyphone representation vector relative to the phrase-level feature representation vector, multiplies the attention value by the polyphone representation vector to obtain a polyphone representation vector matrix, and performs matrix addition or matrix multiplication on the polyphone representation vector and the phrase-level feature representation vector to obtain the attention vector.

204: Splice the attention vector and the polyphone representation vector to obtain the target vector.

After obtaining the attention vector, the server performs matrix multiplication or matrix addition on the attention vector and the polyphone representation vector to obtain the target vector; or the server performs weighted summation on the attention vector and the polyphone representation vector to obtain the target vector. The target vector is obtained through the preset feed-forward attention mechanism, and indicates which phrase in the to-be-processed Chinese sentence is more important and needs a higher weight for the target polyphone, thus improving accuracy of contextual semantic fusion of the target polyphone.

205: Calculate a target pinyin probability of the target vector through a preset linear layer, and determine a target pronunciation of the target polyphone based on the target pinyin probability.

Specifically, the server calculates a probability of the target vector based on each pinyin through a preset linear layer to obtain a polyphone pinyin probability value set; sorts polyphone pinyin probability values in the polyphone pinyin probability value set in descending order of values, and determines the polyphone pinyin probability value ranked first as a target pinyin probability; and determines a pinyin corresponding to the target pinyin probability as the target pronunciation of the target polyphone.

For example, there is one linear layer. The server inputs the target vector into the preset linear layer, and calculates the probability of the target vector based on each pinyin through the linear layer to obtain a polyphone pinyin probability value set including polyphone pinyin probability value 1 and polyphone pinyin probability value 2, and sorts polyphone pinyin probability value 1 and polyphone pinyin probability value 2 in descending order of values to obtain a sequence "polyphone pinyin probability value 2-polyphone pinyin probability value 1". Polyphone pinyin probability value 2 is ranked first, which is used as the target pinyin probability, and the server determines a pinyin corresponding to the target pinyin probability as the target pronunciation of the target polyphone.

Specifically, after calculating the target pinyin probability of the target vector through the preset linear layer and determining the target pronunciation of the target polyphone based on the target pinyin probability, the server obtains an error value of the target pronunciation based on a labeled pronunciation, and optimizes a target pronunciation obtaining policy based on the error value, where the obtaining policy includes an execution process, an algorithm and a network structure for obtaining the target pronunciation.

The server obtains the labeled pronunciation of the target polyphone, and the labeled pronunciation is a pronunciation of the target polyphone based on a sentence corresponding to semantics and an emotion of the to-be-processed Chinese sentence. The labeled pronunciation may be obtained by manual labeling or by labeling through a pre-trained polyphone labeling model. The server calculates a pronunciation similarity between the target pronunciation and the labeled pronunciation of the target polyphone, calculates a difference between the pronunciation similarity and 1 to obtain an error value of the target pronunciation based on the labeled pronunciation, and uses the error value to adjust the execution process of obtaining the target pronunciation and optimize a network structure used for obtaining the target pronunciation. The network structure includes a neural network structure and model parameters, and corresponding processing functions may be representation vector generation, representation vector extraction, phrase segmentation and pinyin probability calculation at a linear layer, etc. The server uses the

error value to add or delete an algorithm for obtaining the target pronunciation or adjusts an execution sequence, and optimizes the target pronunciation obtaining policy based on the error value to improve accuracy of predicting a pronunciation of a polyphone.

In the embodiment of the present application, a character representation vector set is converted into a phrase-level feature representation vector based on a target segmented phrase, and character features are converted to phrase-level features, so as to avoid the out-of-vocabulary problem, thereby effectively improving accuracy of predicting a pronunciation of a polyphone. Splicing processing based on an attention mechanism is performed on a polyphone representation vector and the phrase-level feature representation vector. A target pinyin probability of a target vector is calculated through a preset linear layer. A target pronunciation of a target polyphone is determined based on the target pinyin probability. A pronunciation of the target polyphone is predicted by combining the target segmented phrase and the attention mechanism. Such practice reduces impact of the labeling error problem existing in segmented phrases without any rule or artificial feature design, can accurately capture text semantic information of a to-be-processed Chinese sentence, and improves the accuracy of predicting a pronunciation of a polyphone.

The method for predicting a pronunciation of a polyphone according to the embodiment of the present application is described above. An apparatus for predicting a pronunciation of a polyphone according to an embodiment of the present application is described below. Referring to FIG. 3, an apparatus for predicting a pronunciation of a polyphone in an embodiment of the present application includes:

an obtaining module 301, configured to obtain a labeled to-be-processed Chinese sentence, and obtain a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone;

a conversion module 302, configured to perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and convert the character representation vector set into a phrase-level feature representation vector based on the target

segmented phrases;

a splicing module 303, configured to perform splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

a determining module 304, configured to calculate a target pinyin probability of the target vector through a preset linear layer, and determine a target pronunciation of the target polyphone based on the target pinyin probability.

The function implementation of each module in the foregoing apparatus for predicting a pronunciation of a polyphone corresponds to each step in the foregoing embodiment of the method for predicting a pronunciation of a polyphone, and functions and an implementation process thereof are not described in detail herein.

In the embodiment of the present application, a character representation vector set is converted into a phrase-level feature representation vector based on a target segmented phrase, and character features are converted to phrase-level features, so as to avoid the out-of-vocabulary problem, thereby effectively improving accuracy of predicting a pronunciation of a polyphone. Splicing processing based on an attention mechanism is performed on a polyphone representation vector and the phrase-level feature representation vector. A target pinyin probability of a target vector is calculated through a preset linear layer. A target pronunciation of a target polyphone is determined based on the target pinyin probability. A pronunciation of the target polyphone is predicted by combining the target segmented phrase and the attention mechanism. Such practice reduces impact of the labeling error problem existing in segmented phrases without any rule or artificial feature design, can accurately capture text semantic information of a to-be-processed Chinese sentence, and improves the accuracy of predicting a pronunciation of a polyphone.

Referring to FIG. 4, an apparatus for predicting a pronunciation of a polyphone according to another embodiment of the present application includes:

an obtaining module 301, configured to obtain a labeled to-be-processed Chinese sentence, and obtain a character representation vector set and a polyphone representation vector of the

to-be-processed Chinese sentence, where the to-be-processed Chinese sentence includes a target polyphone;

a conversion module 302, configured to perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and convert the character representation vector set into a phrase-level feature representation vector based on the target segmented phrases;

a splicing module 303, configured to perform splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector;

where the splicing module 303 specifically includes:

a calculation unit 3031, configured to perform attention calculation on the polyphone representation vector and the phrase-level feature representation vector through a preset feed-forward attention mechanism to obtain an attention vector; and

a splicing unit 3032, configured to splice the attention vector and the polyphone representation vector to obtain the target vector; and

a determining module 304, configured to calculate a target pinyin probability of the target vector through a preset linear layer, and determine a target pronunciation of the target polyphone based on the target pinyin probability.

Optionally, the conversion module 302 may be further specifically configured to:

perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain the target segmented phrase;

divide the character representation vector set based on the target segmented phrases to obtain a representation vector set of each phrase; and

perform hybrid pooling on the representation vector set of each phrase through a preset hybrid pooling layer to obtain the phrase-level feature representation vector.

Optionally, the determining module 304 may be further specifically configured to:

calculate a probability of the target vector based on each pinyin through the preset linear

layer to obtain a polyphone pinyin probability value set;

sort polyphone pinyin probability values in the polyphone pinyin probability value set in descending order of values, and determine the polyphone pinyin probability value ranked first as a target pinyin probability; and

determine a pinyin corresponding to the target pinyin probability as the target pronunciation of the target polyphone.

Optionally, the obtaining module 301 includes:

an obtaining unit 3011, configured to obtain an initial Chinese sentence, the target polyphone in the initial Chinese sentence and polyphone position information corresponding to the target polyphone;

a labeling unit 3012, configured to label the target polyphone in the initial Chinese sentence based on the polyphone position information to obtain the to-be-processed Chinese sentence; and

an encoding and extraction unit 3013, configured to sequentially perform character vector encoding and polyphone vector extraction on the to-be-processed Chinese sentence to obtain the character representation vector set and the polyphone representation vector.

Optionally, the encoding and extraction unit 3013 may be further specifically configured to:

encode each character in the to-be-processed Chinese sentence by using a preset deep neural network encoder to obtain the character representation vector set, where one character representation vector corresponds to one character; and

extract a representation vector corresponding to the target polyphone from the character representation vector set based on the polyphone position information, so as to obtain the polyphone representation vector.

Optionally, the apparatus for predicting a pronunciation of a polyphone further includes:

an optimization module 305, configured to obtain an error value of the target pronunciation based on a labeled pronunciation, and optimize a target pronunciation obtaining policy based on the error value, where the obtaining policy includes an execution process, an algorithm and a network structure for obtaining the target pronunciation.

The function implementation of each module and each unit in the foregoing apparatus for predicting a pronunciation of a polyphone corresponds to each step in the foregoing embodiment of the method for predicting a pronunciation of a polyphone, and functions and an implementation process thereof are not described in detail herein.

In the embodiment of the present application, a character representation vector set is converted into a phrase-level feature representation vector based on a target segmented phrase, and character features are converted to phrase-level features, so as to avoid the out-of-vocabulary problem, thereby effectively improving accuracy of predicting a pronunciation of a polyphone. Splicing processing based on an attention mechanism is performed on a polyphone representation vector and the phrase-level feature representation vector. A target pinyin probability of a target vector is calculated through a preset linear layer. A target pronunciation of a target polyphone is determined based on the target pinyin probability. A pronunciation of the target polyphone is predicted by combining the target segmented phrase and the attention mechanism. Such practice reduces impact of the labeling error problem existing in segmented phrases without any rule or artificial feature design, can accurately capture text semantic information of a to-be-processed Chinese sentence, and improves the accuracy of predicting a pronunciation of a polyphone.

The apparatus for predicting a pronunciation of a polyphone in the embodiment of the present application is described in detail above from the perspective of modular functional entities with reference to FIG. 3 and FIG. 4. A device for predicting a pronunciation of a polyphone in an embodiment of the present application is described in detail below from the perspective of hardware processing.

FIG. 5 is a schematic structural diagram of a device for predicting a pronunciation of a polyphone according to an embodiment of the present application. The device 500 for predicting a pronunciation of a polyphone may have relatively large differences due to a configuration or performance difference and may include one or more processors (central processing units (CPUs)) 510, a memory 520, and one or more storage media 530 (such as one or more mass storage devices) that store an application program 533 or data 532. The memory 520 and the storage

medium 530 may be used for short-time storage or persistent storage. The program stored in the storage medium 530 may include one or more modules (not shown in the figure), and each module may include a series of operation instructions in the device 500 for predicting a pronunciation of a polyphone. Further, the processor 510 may be configured to communicate with the storage medium 530, and execute a series of operation instructions in the storage medium 530 on the device 500 for predicting a pronunciation of a polyphone.

The device 500 for predicting a pronunciation of a polyphone may further include one or more power supplies 540, one or more wired or wireless network interfaces 550, one or more input/output interfaces 560, and/or one or more operating systems 531, such as Windows Server, Mac OS X, Unix, Linux and FreeBSD. It can be understood by a person skilled in the art that the structure of the device for predicting a pronunciation of a polyphone shown in FIG. 5 does not constitute a limitation to the device for predicting a pronunciation of a polyphone, and may include more or fewer components than those shown, or combine some components, or have different component arrangements.

The present application further provides a computer-readable storage medium. The computer-readable storage medium may be a non-volatile computer-readable storage medium or a volatile computer-readable storage medium. The computer-readable storage medium stores instructions, and when the instructions are run on a computer, the computer is enabled to perform the steps of the method for predicting a pronunciation of a polyphone.

Further, the computer-readable storage medium may mainly include a program storage area and a data storage area. The program storage area may store an operating system, at least one application program required for functions, etc.; and the data storage area may store data created based on use of blockchain nodes, etc.

The blockchain described in the present application is a new application mode of computer technologies such as distributed data storage, point-to-point transmission, a consensus mechanism and an encryption algorithm. The blockchain is essentially a decentralized database, and is a series of data blocks associated using a cryptographic method. Each data block contains

information about a batch of network transactions, which is used to verify the validity of the information (anti-counterfeiting) and generate the next block. The blockchain may include a blockchain underlying platform, a platform product service layer, an application service layer, etc.

A person skilled in the art can clearly understand that for ease and brevity of description, for specific working processes of the system, apparatus and units described above, reference may be made to the corresponding processes in the foregoing method embodiments. Details are not repeated herein.

The integrated units may be stored in a computer-readable storage medium if implemented in the form of a software function unit and sold or used as a standalone product. Based on such an understanding, the technical solutions of this application essentially, or the part contributing to the prior art, or all or some of the technical solutions may be implemented in the form of a software product. The computer software product is stored in a storage medium and includes several instructions for instructing a computer device (which may be a personal computer, a server, or a network device) to perform all or some of the steps of the methods described in embodiments of the present application. The foregoing storage media include: a USB flash disk, a mobile hard disk, a read-only memory (ROM), a random access memory (RAM), a magnetic disk or an optical disc, and other media that can store program code.

The foregoing embodiments are only used to illustrate the technical solutions of the present application, rather than constitute a limitation thereto. Although the present application is described in detail with reference to the foregoing embodiments, it should be understood by a person of ordinary skill in the art that he/she may still modify the technical solutions described in the foregoing embodiments or equivalently replace some technical features therein; and these modifications or replacements do not make the essence of the corresponding technical solutions depart from the spirit and scope of the technical solutions of various embodiments of the present application.

CLAIMS

1. A method for predicting a pronunciation of a polyphone, comprising:

obtaining a labeled to-be-processed Chinese sentence, and obtaining a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, wherein the to-be-processed Chinese sentence comprises a target polyphone;

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability.

2. The method for predicting a pronunciation of a polyphone according to claim 1, wherein the performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrases comprises:

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain the target segmented phrase;

dividing the character representation vector set based on the target segmented phrase to obtain a representation vector set of each phrase; and

performing hybrid pooling on the representation vector set of each phrase through a preset hybrid pooling layer to obtain the phrase-level feature representation vector.

3. The method for predicting a pronunciation of a polyphone according to claim 1, wherein the performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector comprises:

performing attention calculation on the polyphone representation vector and the phrase-level feature representation vector through a preset feed-forward attention mechanism to obtain an attention vector; and

splicing the attention vector and the polyphone representation vector to obtain the target vector.

4. The method for predicting a pronunciation of a polyphone according to claim 1, wherein the calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability comprises:

calculating a probability of the target vector based on each pinyin through the preset linear layer to obtain a polyphone pinyin probability value set;

sorting polyphone pinyin probability values in the polyphone pinyin probability value set in descending order of values, and determining the polyphone pinyin probability value ranked first as a target pinyin probability; and

determining a pinyin corresponding to the target pinyin probability as the target pronunciation of the target polyphone.

5. The method for predicting a pronunciation of a polyphone according to claim 1, wherein the obtaining a labeled to-be-processed Chinese sentence, and obtaining a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, wherein the to-be-processed Chinese sentence comprises a target polyphone comprises:

obtaining an initial Chinese sentence, the target polyphone in the initial Chinese sentence and polyphone position information corresponding to the target polyphone;

labeling the target polyphone in the initial Chinese sentence based on the polyphone position information to obtain the to-be-processed Chinese sentence; and

sequentially performing character vector encoding and polyphone vector extraction on the to-be-processed Chinese sentence to obtain the character representation vector set and the polyphone representation vector.

6. The method for predicting a pronunciation of a polyphone according to claim 5, wherein the sequentially performing character vector encoding and polyphone vector extraction on the to-be-processed Chinese sentence to obtain the character representation vector set and the polyphone representation vector comprises:

encoding each character in the to-be-processed Chinese sentence by using a preset deep neural network encoder to obtain the character representation vector set, wherein one character representation vector corresponds to one character; and

extracting a representation vector corresponding to the target polyphone from the character representation vector set based on the polyphone position information, so as to obtain the polyphone representation vector.

7. The method for predicting a pronunciation of a polyphone according to any one of claims 1 to 6, wherein after the calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability, the method further comprises:

obtaining an error value of the target pronunciation based on a labeled pronunciation, and optimizing a target pronunciation obtaining policy based on the error value, wherein the obtaining policy comprises an execution process, an algorithm and a network structure for obtaining the target pronunciation.

8. An apparatus for predicting a pronunciation of a polyphone, comprising:

an obtaining module, configured to obtain a labeled to-be-processed Chinese sentence, and obtain a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, wherein the to-be-processed Chinese sentence comprises a target polyphone;

a conversion module, configured to perform phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and convert the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

a splicing module, configured to perform splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

a determining module, configured to calculate a target pinyin probability of the target vector through a preset linear layer, and determine a target pronunciation of the target polyphone based on the target pinyin probability.

9. A device for predicting a pronunciation of a polyphone, comprising a memory and at least one processor, wherein the memory store instructions; and

the at least one processor invokes the instructions in the memory to enable the device for predicting a pronunciation of a polyphone to perform the method for predicting a pronunciation of a polyphone comprising the following steps:

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability.

10. The device for predicting a pronunciation of a polyphone according to claim 9, wherein when the device for predicting a pronunciation of a polyphone performs, by the processor, the steps of performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrases, the following steps are comprised:

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain the target segmented phrase;

dividing the character representation vector set based on the target segmented phrase to

obtain a representation vector set of each phrase; and

performing hybrid pooling on the representation vector set of each phrase through a preset hybrid pooling layer to obtain the phrase-level feature representation vector.

11. The device for predicting a pronunciation of a polyphone according to claim 9, wherein when the device for predicting a pronunciation of a polyphone performs, by using the processor, the steps of performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector, the following steps are comprised:

performing attention calculation on the polyphone representation vector and the phrase-level feature representation vector through a preset feed-forward attention mechanism to obtain an attention vector; and

splicing the attention vector and the polyphone representation vector to obtain the target vector.

12. The device for predicting a pronunciation of a polyphone according to claim 9, wherein when the device for predicting a pronunciation of a polyphone performs, by using the processor, the steps of calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability, the following steps are comprised:

calculating a probability of the target vector based on each pinyin through the preset linear layer to obtain a polyphone pinyin probability value set;

sorting polyphone pinyin probability values in the polyphone pinyin probability value set in descending order of values, and determining the polyphone pinyin probability value ranked first as a target pinyin probability; and

determining a pinyin corresponding to the target pinyin probability as the target pronunciation of the target polyphone.

13. The device for predicting a pronunciation of a polyphone according to claim 9, wherein when the device for predicting a pronunciation of a polyphone performs, by using the processor,

the steps of obtaining a labeled to-be-processed Chinese sentence, and obtaining a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, wherein the to-be-processed Chinese sentence comprises a target polyphone, the following steps are comprised:

obtaining an initial Chinese sentence, the target polyphone in the initial Chinese sentence and polyphone position information corresponding to the target polyphone;

labeling the target polyphone in the initial Chinese sentence based on the polyphone position information to obtain the to-be-processed Chinese sentence; and

sequentially performing character vector encoding and polyphone vector extraction on the to-be-processed Chinese sentence to obtain the character representation vector set and the polyphone representation vector.

14. The device for predicting a pronunciation of a polyphone according to claim 13, wherein when the device for predicting a pronunciation of a polyphone performs, by using the processor, the steps of sequentially performing character vector encoding and polyphone vector extraction on the to-be-processed Chinese sentence to obtain the character representation vector set and the polyphone representation vector, the following steps are comprised:

encoding each character in the to-be-processed Chinese sentence by using a preset deep neural network encoder to obtain the character representation vector set, wherein one character representation vector corresponds to one character; and

extracting a representation vector corresponding to the target polyphone from the character representation vector set based on the polyphone position information, so as to obtain the polyphone representation vector.

15. The device for predicting a pronunciation of a polyphone according to any one of claims 9 to 14, wherein after the device for predicting a pronunciation of a polyphone performs, by using the processor, the steps of calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability, the following steps are further comprised:

obtaining an error value of the target pronunciation based on a labeled pronunciation, and optimizing a target pronunciation obtaining policy based on the error value, wherein the obtaining policy comprises an execution process, an algorithm and a network structure for obtaining the target pronunciation.

16. A computer-readable storage medium, wherein the computer-readable storage medium stores instructions, and when the instructions are executed by a processor, the method for predicting a pronunciation of a polyphone is implemented as follows:

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrase;

performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector; and

calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability.

17. The computer-readable storage medium according to claim 16, wherein when the instructions for predicting a pronunciation of a polyphone are executed by the processor to perform the steps of performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain a target segmented phrase, and converting the character representation vector set into a phrase-level feature representation vector based on the target segmented phrases, the following steps are comprised:

performing phrase segmentation processing on the to-be-processed Chinese sentence to obtain the target segmented phrase;

dividing the character representation vector set based on the target segmented phrase to obtain a representation vector set of each phrase; and

performing hybrid pooling on the representation vector set of each phrase through a preset hybrid pooling layer to obtain the phrase-level feature representation vector.

18. The computer-readable storage medium according to claim 16, wherein when the

instructions for predicting a pronunciation of a polyphone are executed by the processor to perform the steps of performing splicing processing on the polyphone representation vector and the phrase-level feature representation vector based on an attention mechanism to obtain a target vector, the following steps are comprised:

performing attention calculation on the polyphone representation vector and the phrase-level feature representation vector through a preset feed-forward attention mechanism to obtain an attention vector; and

splicing the attention vector and the polyphone representation vector to obtain the target vector.

19. The computer-readable storage medium according to claim 16, wherein when the instructions for predicting a pronunciation of a polyphone are executed by the processor to perform the steps of calculating a target pinyin probability of the target vector through a preset linear layer, and determining a target pronunciation of the target polyphone based on the target pinyin probability, the following steps are comprised:

calculating a probability of the target vector based on each pinyin through the preset linear layer to obtain a polyphone pinyin probability value set;

sorting polyphone pinyin probability values in the polyphone pinyin probability value set in descending order of values, and determining the polyphone pinyin probability value ranked first as a target pinyin probability; and

determining a pinyin corresponding to the target pinyin probability as the target pronunciation of the target polyphone.

20. The computer-readable storage medium according to claim 16, wherein when the instructions for predicting a pronunciation of a polyphone are executed by the processor to perform the steps of obtaining a labeled to-be-processed Chinese sentence, and obtaining a character representation vector set and a polyphone representation vector of the to-be-processed Chinese sentence, wherein the to-be-processed Chinese sentence comprises a target polyphone, the following steps are comprised:

obtaining an initial Chinese sentence, the target polyphone in the initial Chinese sentence and polyphone position information corresponding to the target polyphone;

labeling the target polyphone in the initial Chinese sentence based on the polyphone position information to obtain the to-be-processed Chinese sentence; and

sequentially performing character vector encoding and polyphone vector extraction on the to-be-processed Chinese sentence to obtain the character representation vector set and the polyphone representation vector.