

WORD VECTOR PROCESSING METHOD AND APPARATUS

Claim of Priority

This application claims priority to Chinese Patent Application No. 201710045459.7 filed on January 22, 2017, the entire contents of which are hereby
5 incorporated by reference.

Technical Field

The present application relates to the technical field of computer software technologies, and in particular, to a word vector processing method and apparatus.

Background Art

10 Current natural language processing solutions mostly use a neural network-based architecture. An important basic technology in such an architecture is a word vector. The word vector is a vector for mapping a word to a fixed dimension, and represents semantic information of the word.

In the prior art, common word vector generation algorithms are designed specially
15 for English, for example, Google's word vector algorithm, Facebook's n-gram character algorithm, Microsoft's deep neural network algorithm, and the like.

However, these algorithms in the prior art are inapplicable to Chinese, or although the algorithms are applicable to Chinese, word vectors of Chinese words generated thereby have a poor practical effect.

20 Summary of the Invention

Embodiments of the present application provide a word vector processing method and apparatus, so as to solve a problem that word vector generation algorithms in the prior art are inapplicable to Chinese, or although the algorithms are applicable to Chinese, word vectors of Chinese words generated thereby have a poor practical effect.

To solve the foregoing technical problem, the embodiments of the present application are implemented as follows:

The embodiments of the present application provide a word vector processing method, including:

- 5 performing word segmentation on a corpus to obtain words;
- determining n-gram strokes corresponding to the words, the n-gram stroke representing n successive strokes of a corresponding word;
- establishing and initializing word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words; and
- 10 training the word vectors and the stroke vectors according to the corpus obtained after the word segmentation, the word vectors, and the stroke vectors.

The embodiments of the present application provide a word vector processing apparatus, including:

- a word segmentation module configured to perform word segmentation on a corpus
- 15 to obtain words;
- a determining module configured to determine n-gram strokes corresponding to the words, the n-gram stroke representing n successive strokes of a corresponding word;
- an initialization module configured to establish and initialize word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words; and
- 20 a training module configured to train the word vectors and the stroke vectors according to the corpus obtained after the word segmentation, the word vectors, and the stroke vectors.

By using at least one of the foregoing technical solutions, the embodiments of the present application can achieve the following beneficial effects: features of a word can

25 be shown more precisely by using n-gram strokes corresponding to the word, thus

enhancing accuracy of word vectors of Chinese words, and achieving a desirable practical effect. Therefore, the present application can solve some or all problems in the prior art.

Brief Description of the Drawings

5 To describe the technical solutions in the embodiments of the present application or in the prior art more clearly, the following briefly introduces the accompanying drawings required for describing the embodiments or the prior art. Apparently, the accompanying drawings in the following description show merely some embodiments of the present invention, and those of ordinary skill in the art may still derive other drawings from
10 these accompanying drawings without creative efforts.

FIG. 1 is a schematic flowchart of a word vector processing method according to an embodiment of the present application;

FIG. 2 is a schematic flowchart of a specific implementation solution of the word vector processing method in an actual application scenario according to an embodiment
15 of the present application;

FIG. 3 is a schematic diagram of a related processing action on a part of the corpus used in the process of FIG. 2 according to an embodiment of the present application;
and

FIG. 4 is a schematic structural diagram of a word vector processing apparatus
20 corresponding to FIG. 1 according to an embodiment of the present application.

Detailed Description

Embodiments of the present application provide a word vector processing method and apparatus.

To make those skilled in the art better understand the technical solution of the present
25 application, the technical solution of the embodiments of the present application is

clearly and completely described below with reference to the accompanying drawings in the embodiments of the present application. Apparently, the described embodiments are some rather than all of the embodiments of the present application. Based on the described embodiments of the present application, all other embodiments obtained by those of ordinary skill in the art without creative effort belong to the protection scope of the present application.

The solution of the present application is applicable to a word vector of a Chinese word, and also to a word vector of a word in another language similar to Chinese, for example, a word vector of a word in a language that features strokes, such as Japanese.

FIG. 1 is a schematic flowchart of a word vector processing method according to an embodiment of the present application. This process may be executed by a program having a word vector generation function and/or a training function. From the device perspective, an entity executing the process may include, but is not limited to, the following devices that can carry the program: a personal computer, a medium/large-scale computer, a computer cluster, a mobile phone, a tablet computer, an intelligent wearable device, a vehicle-mounted product, and the like.

The process of FIG. 1 may include the following steps:

S101: Word segmentation is performed on a corpus to obtain words.

In the embodiment of the present application, the words may specifically be at least some of words that occur at least once in the corpus. For ease of subsequent processing, the words may be stored in a glossary and are read from the glossary when required.

S102: n-gram strokes corresponding to the words are determined, the n-gram stroke representing n successive strokes of a corresponding word.

For ease of understanding, the "n-gram stroke" is further explained by using Chinese as an example. For a Chinese character or word, strokes include "horizontal", "vertical",

"throw away", "press down", "break", "dot", "rise", and the like. The n-gram stroke may represent n successive strokes of one Chinese character or word.

For example, for the character "珍" (which is a Chinese character meaning treasure), corresponding 3-gram strokes thereof are "horizontal, horizontal, and vertical" (the first stroke to the third stroke), "horizontal, vertical, and rise" (the second stroke to the fourth stroke), "vertical, rise, and throw away" (the third stroke to the fifth stroke), and so on; and corresponding 4-gram strokes are "horizontal, horizontal, vertical, and rise" (the first stroke to the fourth stroke), "horizontal, vertical, rise, and throw away" (the second stroke to the fifth stroke), and so on.

For another example, for the term "大人" (which is a Chinese term meaning adult), corresponding 3-gram strokes thereof are "horizontal, throw away, and press down" (the first stroke to the third stroke), "throw away, press down, and throw away" (the second stroke to the fourth stroke), and so on; and corresponding 4-gram strokes are "horizontal, throw away, press down, and throw away" (the first stroke to the fourth stroke), "throw away, press down, throw away, and press down" (the second stroke to the fifth stroke), and so on.

In the embodiment of the present application, a value of n may be dynamically adjusted. For a same word, during determining of n-gram strokes corresponding to the word, n may have only one value (for example, only 3-gram strokes corresponding to the word are determined), or may have multiple values (for example, 3-gram strokes and 4-gram strokes corresponding to the word are determined). When the value of n is a specific value, an n-gram stroke is probably a radical of a character exactly. When the value of n is the total number of strokes of a character or a term, an n-gram stroke is exactly the character or the term.

In the embodiment of the present application, for ease of computer processing, the n-gram stroke may be indicated with a designated code. For example, different strokes are indicated by different numbers respectively, and then the n-gram stroke may be correspondingly represented as a numeric string.

5 S103: Word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words are established and initialized.

In the embodiment of the present application, to ensure an effect of the solution, there are probably some limitations when the word vectors and the stroke vectors are initialized. For example, it is not allowed to initialize the word vectors and the stroke
10 vectors to identical vectors. For another example, values of vector elements of some word vectors or stroke vectors should not all equal 0.

In the embodiment of the present application, the word vectors of the words and the stroke vectors of the n-gram strokes corresponding to the words may be initialized in a random initialization manner or in a manner of initializing according to a specified
15 probability distribution. Stroke vectors of the same n-gram strokes are also the same. The specified probability distribution may be, for example, distribution within 0 to 1.

In addition, if word vectors and stroke vectors corresponding to some words are already trained before on the basis of other corpora, when the word vectors and the stroke vectors corresponding to these words are further trained on the basis of the corpus
20 shown in FIG. 1, the word vectors and the stroke vectors corresponding to these words may not be established and initialized again, but are retrained on the basis of the corpus shown in FIG. 1 and a previous training result.

S104: The word vectors and the stroke vectors are trained according to the corpus obtained after the word segmentation, the word vectors, and the stroke vectors.

In the embodiment of the present application, the training may be implemented by using a neural network. The neural network includes, but is not limited to, a shallow neural network and a deep neural network.

With the method of FIG. 1, features of a word can be shown more precisely by using
5 n-gram strokes corresponding to the word, thus enhancing accuracy of word vectors of Chinese words, and achieving a desirable practical effect. Therefore, the method can solve some or all problems in the prior art.

Based on the method of FIG. 1, an embodiment of the present application further provides some specific implementation solutions and extension solutions of the method,
10 which are described below.

In the embodiment of the present application, step S102 of determining n-gram strokes corresponding to the words may specifically include: determining, according to a result of the word segmentation on the corpus, words occurring at least once in the corpus; and

15 performing the following operation on each determined word:

determining n-gram strokes corresponding to the word, where each n-gram stroke corresponding to the word represents n successive strokes of the word, and n is one positive integer or multiple different positive integers.

Further, if a word occurs rarely in the corpus, corresponding training samples and
20 the number of trainings are also rather few during training based on the corpus, thus adversely affecting reliability of a training result. Therefore, such words may be eliminated and are not trained temporarily, and may be trained in another corpus subsequently.

Based on this thought, the step of determining, according to a result of the word
25 segmentation, words occurring at least once in the corpus may specifically include:

determining, according to the result of the word segmentation on the corpus, words that occur in the corpus for not less than a set number of times, the set number of times being not less than 1. A specific value of the set number of times may be determined according to actual status.

5 In the embodiment of the present application, for step S104, there may be multiple specific training manners, for example, a training manner based on context words, a training manner based on a designated near-synonym or synonym, and the like. For ease of understanding, a detailed description is made by using the former manner as an example.

10 The step of training the word vectors and the stroke vectors according to the corpus obtained after the word segmentation, the word vectors, and the stroke vectors may specifically include: determining a designated word in the corpus obtained after the word segmentation, and one or more context words of the designated word in the corpus obtained after the word segmentation; determining a degree of similarity between the
15 designated word and the context word according to stroke vectors of n-gram strokes corresponding to the designated word as well as a word vector of the context word; and updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the degree of similarity between the designated word and the context word.

20 This application does not limit a specific manner of determining the degree of similarity. For example, the degree of similarity may be calculated on the basis of a cosine operation on an angle between vectors, or on the basis of a quadratic sum operation on vectors, or the like.

There may be multiple designated words. The designated words may repeatedly
25 occur in different positions of the corpus. The processing action described in the

previous paragraph may be performed on each designated word. Preferably, each word in the corpus obtained after the word segmentation may be used as a designated word.

In the embodiment of the present application, the training in step S104 may increase the degree of similarity between the designated word and the context word (herein, the degree of similarity may reflect a degree of association, where a degree of association
5 between a word and its context word is relatively high, and context words corresponding to synonyms or near-synonyms also have the same or similar meanings), and reduce a degree of similarity between the designated word and a non-context word. The non-context word may be used as a negative sample word described below, and then the
10 context word may be used as a positive sample word.

It can be seen from the above that, some negative sample words need to be determined for comparison in a training process. One or more words may be randomly selected as negative sample words from the corpus obtained after the word segmentation, or non-context words may be strictly selected as negative sample words. Using the
15 former manner as an example, the step of updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the degree of similarity between the designated word and the context word may specifically include: selecting one or more words from the words as a negative
20 negative sample word; determining a degree of similarity between the designated word and each negative sample word; determining a loss characterization value corresponding to the designated word according to a designated loss function, the degree of similarity between the designated word and the context word, and the degree of similarity between the designated word and each negative sample word; and updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the
25 designated word according to the loss characterization value.

The loss characterization value is used to measure a degree of an error between a current vector value and a training target. The foregoing several degrees of similarity may be used as parameters of the loss function. A specific expression of the loss function is not limited in the present application, and will be described in detail with an example
5 later.

In the embodiment of the present application, the updating of the word vector and the stroke vectors is actually a modification to the degree of error. When the solution of the present application is implemented by using a neural network, such a modification may be implemented on the basis of counter propagation and gradient descent. In this
10 case, the gradient is a gradient corresponding to the loss function.

The step of updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value may specifically include: determining a gradient corresponding to the loss function according to the loss characterization value; and updating the word
15 vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the gradient.

In the embodiment of the present application, the process of training the word vectors and the stroke vectors may be carried out iteratively on the basis of at least some words in the corpus obtained after the word segmentation, so that the word vectors and
20 the stroke vectors gradually converge till the training is finished.

Training on all words in the corpus obtained after the word segmentation is used as an example. Step S104 of training the word vectors and the stroke vectors according to the corpus obtained after the word segmentation, the word vectors, and the stroke vectors may specifically include:

traversing the corpus obtained after the word segmentation, and performing the following operation on each word in the corpus obtained after the word segmentation:

determining one or more context words of the word in the corpus obtained after the word segmentation; and

5 performing the following operation according to each context word:

determining a degree of similarity between the word and the context word according to stroke vectors of n-gram strokes corresponding to the word and a word vector of the context word; and

10 updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the word according to the degree of similarity between the word and the context word.

A specific update process has been described above, and is not described in detail herein again.

15 Further, for ease of computer processing, the foregoing traversing process may be implemented based on a window.

For example, the step of determining one or more context words of the word in the corpus obtained after the word segmentation may specifically include: with this word as a center, sliding leftward and/or rightward by a distance of a designated number of words in the corpus obtained after the word segmentation, to establish a window; and
20 determining words excluding this word in the window as the context words of this word.

Definitely, it is also feasible to establish a window of a set length starting from the first word in the corpus obtained after the word segmentation, the window including the first word and a set number of successive words following the first word. After the words in the window are processed, the window is slid forward to process a next batch of words
25 in the corpus, till the corpus is completely traversed.

The above describes a word vector processing method according to an embodiment of the present application. For ease of understanding, based on the foregoing description, an embodiment of the present application further provides a schematic flowchart of a specific implementation solution of the word vector processing method in an actual application scenario, as shown in FIG. 2.

A process in FIG. 2 mainly includes the following steps:

Step 1: Word segmentation is performed on a Chinese corpus with a word segmentation tool, a Chinese corpus obtained after the word segmentation is scanned, all words that have occurred are counted to establish a glossary, and words that occur for less than b times (that is, the set number of times) are deleted. Then step 2 is performed.

Step 2: The glossary is scanned one by one, and n -gram strokes corresponding to each word are extracted to establish an n -gram stroke table and a table of mapping between words and corresponding n -gram strokes. A total of five types of strokes are defined: "horizontal", "vertical", "throw away", "press down", and "break" (the strokes "hook", "dot", and "rise" are all regarded as "break"), which are indicated by numbers 1 to 5 respectively. Then step 3 is performed.

Step 3: A d -dimensional word vector is established for each word in the glossary, a d -dimensional stroke vector is also established for each n -gram stroke in the n -gram stroke table, and all the established vectors are randomly initialized. Then step 4 is performed.

Step 4: Slide over the words are one by one starting from the first word in the Chinese corpus obtained after the word segmentation; and one word is selected each time as a "current word w " (that is, the designated word). The process ends if all the words in the whole corpus have been traversed as w , otherwise step 5 is performed.

Step 5: With the current word as a center, slide towards two sides by a length of k words, to establish a window; and one word is selected as a "context word c " each time from the first word to the last word (excluding the current word w) in the window. Step 4 is performed if all the words in the window have been traversed as c , otherwise step 6 is performed.

Step 6: For the current word w , n-gram strokes corresponding to the current word w are searched for in the table of mapping between words and corresponding n-gram strokes in step 2, and a degree of similarity between the current word w and the context word c is calculated by using the following equation (1):

$$sim(w, c) = \sum_{q \in S(w)} \vec{q} \cdot \vec{c} \quad \text{Equation (1)}$$

In this equation, S represents the n-gram stroke table established in step 2, $S(w)$ represents a set of n-gram strokes corresponding to the current word w in the mapping table in step 2, q represents an element (that is, an n-gram stroke) in the set $S(w)$, $sim(w, c)$ represents a score of a degree of similarity between the current word w and the context word c , and $\vec{q} \cdot \vec{c}$ represents a vector dot product operation between the n-gram stroke q and the context word c . Then step 7 is performed.

Step 7: λ words are randomly extracted as negative sample words, and a loss score $l(w, c)$ is calculated according to the following equation (2) (that is, the loss function described above). The loss score may be used as the loss characterization value described above:

$$l(w, c) = \log \sigma(sim(w, c)) + \sum_{i=1}^{\lambda} E_{c' \in p(V)} [\log \sigma(-sim(w, c'))] \quad \text{Equation (2)}$$

In this equation, $\log$ represents a log function, c' represents randomly extracted negative sample words, $E_{c' \in p(V)}[x]$ represents an expected value of the expression x in

the case where the randomly extracted negative sample words c' meet a probability distribution $p(V)$, and $\sigma(\cdot)$ is an excitation function of a neural network. Refer to the following equation (3) for details.

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad \text{Equation (3)}$$

5 In this equation, if x is a real number, $\sigma(x)$ is also a real number. A gradient is calculated according to a value of $l(w, c)$, and a vector $\vec{w}_q$ of an n-gram stroke and a vector $\vec{c}$ of a context word are updated. Then step 5 is performed.

Step 6 and step 7 are key steps in the foregoing steps 1 to 7. For ease of understanding, these two steps are described by using an example with reference to FIG. 3.

10 FIG. 3 is a schematic diagram of a related processing action on a part of the corpus used in the process of FIG. 2 according to an embodiment of the present application.

As shown in FIG. 3, assuming that the corpus includes a sentence "治理雾霾刻不容缓" (which means "smog treatment is of great urgency"), three terms "治理" (which means "treatment"), "雾霾" (which means "smog"), and "刻不容缓" (which means "is_of_great_urgency") in the sentence are obtained after the word segmentation.

It is assumed that "雾霾" is selected as a current word w and "治理" is selected as a context word c . All n-gram strokes $S(w)$ mapped by the current word w are extracted. For example, 3-gram strokes mapped by "雾霾" include "145", "452", "524", and so on. Then, a loss score $l(w, c)$ is calculated according to the equations (1), (2), and (3), and
20 then a gradient is calculated, to update a word vector of c and all stroke vectors corresponding to w .

The above describes a word vector processing method according to an embodiment of the present application. Based on the same concept, an embodiment of the present application further provides a corresponding apparatus, as shown in FIG. 4.

FIG. 4 is a schematic structural diagram of a word vector processing apparatus
5 corresponding to FIG. 1 according to an embodiment of the present application. The apparatus may be located in an entity executing the process of FIG. 1, and includes:

a word segmentation module 401 configured to perform word segmentation on a corpus to obtain words;

a determining module 402 configured to determine n-gram strokes corresponding to
10 the words, the n-gram stroke representing n successive strokes of a corresponding word;

an initialization module 403 configured to establish and initialize word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words; and

a training module 404 configured to train the word vectors and the stroke vectors according to the corpus obtained after the word segmentation, the word vectors, and the
15 stroke vectors.

Optionally, the determining, by the determining module 402, n-gram strokes corresponding to the words specifically includes:

determining, by the determining module 402 according to a result of the word segmentation on the corpus, words occurring at least once in the corpus; and

20 performing the following operation on each determined word:

determining n-gram strokes corresponding to the word, where each n-gram stroke corresponding to the word represents n successive strokes of the word, and n is one positive integer or multiple different positive integers.

Optionally, the determining, by the determining module 402 according to a result of
25 the word segmentation, words occurring at least once in the corpus specifically includes:

determining, by the determining module 402 according to the result of the word segmentation on the corpus, a word that occurs in the corpus for not less than a set number of times, the set number of times being not less than 1.

Optionally, the initializing, by the initialization module 403, word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words specifically includes:

initializing, by the initialization module 403, the word vectors of the words and the stroke vectors of the n-gram strokes corresponding to the words in a random initialization manner or in a manner of initializing according to a specified probability distribution, where stroke vectors of the same n-gram strokes are also the same.

Optionally, the training, by the training module 404, the word vectors and the stroke vectors according to the corpus obtained after the word segmentation, the word vectors, and the stroke vectors specifically includes:

determining, by the training module 404, a designated word in the corpus obtained after the word segmentation, and one or more context words of the designated word in the corpus obtained after the word segmentation;

determining a degree of similarity between the designated word and the context word according to stroke vectors of n-gram strokes corresponding to the designated word as well as a word vector of the context word; and

updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the degree of similarity between the designated word and the context word.

Optionally, the updating, by the training module 404, the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word

according to the degree of similarity between the designated word and the context word specifically includes:

selecting, by the training module 404, one or more words from the words as a negative sample word;

5 determining a degree of similarity between the designated word and each negative sample word;

determining a loss characterization value corresponding to the designated word according to a designated loss function, the degree of similarity between the designated word and the context word, and the degree of similarity between the designated word
10 and each negative sample word; and

updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value.

Optionally, the updating, by the training module 404, the word vector of the context
15 word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value specifically includes:

determining, by the training module 404, a gradient corresponding to the loss function according to the loss characterization value; and

updating the word vector of the context word and the stroke vectors of the n-gram
20 strokes corresponding to the designated word according to the gradient.

Optionally, the selecting, by the training module 404, one or more words from the words as a negative sample word specifically includes:

randomly selecting, by the training module 404, one or more words from the words as the negative sample word.

Optionally, the training, by the training module 404, the word vectors and the stroke vectors according to the corpus obtained after the word segmentation, the word vectors, and the stroke vectors specifically includes:

traversing, by the training module 404, the corpus obtained after the word
5 segmentation, and performing the following operation on each word in the corpus obtained after the word segmentation:

determining one or more context words of the word in the corpus obtained after the word segmentation; and

performing the following operation according to each context word:

10 determining a degree of similarity between the word and the context word according to stroke vectors of n-gram strokes corresponding to the word and a word vector of the context word; and

updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the word according to the degree of similarity between the
15 word and the context word.

Optionally, the determining, by the training module 404, one or more context words of the word in the corpus obtained after the word segmentation specifically includes:

with this word as a center, sliding, by the training module 404, leftward and/or rightward by a distance of a designated number of words in the corpus obtained after the
20 word segmentation, to establish a window; and

determining words excluding this word in the window as the context words of this word.

Optionally, the words are Chinese words, and the word vectors are word vectors of the Chinese words.

The apparatus provided in the embodiment of the present application corresponds to the method in a one-to-one manner. Therefore, the apparatus also has beneficial technical effects similar to the corresponding method. Because the beneficial technical effects of the method have been described in detail above, the beneficial technical effects of the corresponding apparatus are not described herein again.

In the 1990s, an improvement on a technology may be obviously distinguished as an improvement on hardware (for example, an improvement on a circuit structure such as a diode, a transistor, and a switch) or an improvement on software (an improvement on a method procedure). However, with the development of technologies, improvements of many method procedures at present may be considered as direct improvements on hardware circuit structures. Almost all designers program the improved method procedures into hardware circuits to obtain corresponding hardware circuit structures. Therefore, it is improper to assume that the improvement of a method procedure cannot be implemented by using a hardware entity module. For example, a Programmable Logic Device (PLD) (for example, a Field Programmable Gate Array (FPGA)) is such an integrated circuit whose logic functions are determined by devices programmed by a user. Designers program by themselves to "integrate" a digital system into a piece of PLD, without the need to ask a chip manufacturer to design and manufacture a dedicated integrated circuit chip. Moreover, at present, the programming is mostly implemented by using logic compiler software, instead of manually manufacturing an integrated circuit chip. The logic compiler software is similar to a software compiler used for developing and writing a program, and original code before compiling also needs to be written by using a specific programming language, which is referred to as a Hardware Description Language (HDL). There are many types of HDLs, such as Advanced Boolean Expression Language (ABEL), Altera Hardware Description Language

(AHDL), Confluence, Cornell University Programming Language (CUPL), HDCal, Java Hardware Description Language (JHDL), Lava, Lola, MyHDL, PALASM, and Ruby Hardware Description Language (RHDL), among which Very-High-Speed Integrated Circuit Hardware Description Language (VHDL) and Verilog are most commonly used now. Those skilled in the art also should know that a hardware circuit for implementing the logic method procedure may be easily obtained by slightly logically programming the method procedure using the above several hardware description languages and programming it into an integrated circuit.

A controller may be implemented in any suitable manner. For example, the controller may be in the form of, for example, a microprocessor or a processor and a computer readable medium storing computer readable program code (for example, software or firmware) executable by the (micro)processor, a logic gate, a switch, an Application Specific Integrated Circuit (ASIC), a programmable logic controller, and an embedded micro-controller. Examples of the controller include, but are not limited to, the following micro-controllers: ARC 625D, Atmel AT91SAM, Microchip PIC18F26K20, and Silicone Labs C8051F320. A memory controller may also be implemented as a part of control logic of a memory. Those skilled in the art also know that, the controller may be implemented by using pure computer readable program code, and in addition, the method steps may be logically programmed to enable the controller to implement the same function in a form of a logic gate, a switch, an application specific integrated circuit, a programmable logic controller and an embedded microcontroller. Therefore, this type of controller may be considered as a hardware component, and apparatuses included therein for implementing various functions may also be considered as structures inside the hardware component. Or, the apparatuses used for implementing various functions

may even be considered as both software modules for implementing the method and structures inside the hardware component.

The system, apparatus, module or unit illustrated in the above embodiments may be specifically implemented by using a computer chip or an entity, or a product having a
5 certain function. A typical implementation device is a computer. Specifically, the computer may be, for example, a personal computer, a laptop computer, a cellular phone, a camera phone, a smart phone, a personal digital assistant, a media player, a navigation device, an email device, a game console, a tablet computer, a wearable device, or a combination of any of these devices.

10 For ease of description, when the apparatus is described, it is divided into various units in terms of functions for respective descriptions. Definitely, when the present application is implemented, functions of the units may be implemented in the same or multiple pieces of software and/or hardware.

Those skilled in the art should understand that, the embodiments of the present
15 invention may be provided as a method, a system, or a computer program product. Therefore, the present invention may be implemented as a complete hardware embodiment, a complete software embodiment, or an embodiment combining software and hardware. Moreover, the present invention may be a computer program product implemented on one or more computer usable storage media (including, but not limited
20 to, a magnetic disk memory, a CD-ROM, an optical memory, and the like) including computer usable program code.

The present application is described with reference to flowcharts and/or block diagrams according to the method, device (system) and computer program product according to the embodiments of the present invention. It should be understood that a
25 computer program instruction may be used to implement each process and/or block in

the flowcharts and/or block diagrams and combinations of processes and/or blocks in the flowcharts and/or block diagrams. These computer program instructions may be provided for a general-purpose computer, a special-purpose computer, an embedded processor, or a processor of any other programmable data processing device to generate
5 a machine, so that the instructions executed by a computer or a processor of any other programmable data processing device generate an apparatus for implementing a specified function in one or more processes in the flowcharts and/or in one or more blocks in the block diagrams.

These computer program instructions may also be stored in a computer readable
10 memory that can instruct the computer or any other programmable data processing device to work in a particular manner, such that the instructions stored in the computer readable memory generate an artifact that includes an instruction apparatus. The instruction apparatus implements a specified function in one or more processes in the flowcharts and/or in one or more blocks in the block diagrams.

15 These computer program instructions may also be loaded onto a computer or another programmable data processing device, such that a series of operation steps are performed on the computer or another programmable device, thereby generating computer-implemented processing. Therefore, the instructions executed on the computer or another programmable device provide steps for implementing a specified
20 function in one or more processes in the flowcharts and/or in one or more blocks in the block diagrams.

In a typical configuration, a computation device includes one or more central processing units (CPUs), an I/O interface, a network interface, and a memory.

The memory may include computer readable media such as a volatile memory, a
25 Random Access Memory (RAM), and/or non-volatile memory, e.g., Read-Only

Memory (ROM) or flash RAM. The memory is an example of a computer readable medium.

The computer readable medium includes non-volatile and volatile media as well as movable and non-movable media, and can implement information storage by means of any method or technology. Information may be a computer readable instruction, a data structure, and a module of a program or other data. A storage medium of a computer includes, for example, but is not limited to, a phase change memory (PRAM), a static random access memory (SRAM), a dynamic random access memory (DRAM), other types of RAMs, a ROM, an electrically erasable programmable read-only memory (EEPROM), a flash memory or other memory technologies, a compact disk read-only memory (CD-ROM), a digital versatile disc (DVD) or other optical storages, a cassette tape, a magnetic tape/magnetic disk storage or other magnetic storage devices, or any other non-transmission medium, and can be used to store information accessed by the computing device. According to the definition of this text, the computer readable medium does not include transitory media, such as a modulated data signal and a carrier.

It should be further noted that, the term "include", "comprise" or other variations thereof are intended to cover non-exclusive inclusion, so that a process, method, commodity or device including a series of elements not only includes the elements, but also includes other elements not clearly listed, or further includes inherent elements of the process, method, commodity or device. In a case without any more limitations, an element defined by "including a/an..." does not exclude that the process, method, commodity or device including the element further has other identical elements.

Those skilled in the art should understand that, the embodiments of the present application may be provided as a method, a system, or a computer program product. Therefore, the present application may be implemented as a complete hardware

embodiment, a complete software embodiment, or an embodiment combining software and hardware. Moreover, the present application may be in the form of a computer program product implemented on one or more computer usable storage media (including, but not limited to, a magnetic disk memory, a CD-ROM, an optical memory and the like) including computer usable program code.

The present application may be described in a common context of a computer executable instruction executed by a computer, for example, a program module. Generally, the program module includes a routine, a program, an object, an assembly, a data structure, and the like used for executing a specific task or implementing a specific abstract data type. The present application may also be implemented in distributed computing environments, and in the distributed computer environments, a task is executed by using remote processing devices connected through a communications network. In the distributed computer environment, the program module may be located in local and remote computer storage media including a storage device.

The embodiments in the specification are described progressively, identical or similar parts of the embodiments may be obtained with reference to each other, and each embodiment emphasizes a part different from other embodiments. Especially, the system embodiment is basically similar to the method embodiment, so it is described simply, and for related parts, reference may be made to the descriptions of the parts in the method embodiment.

The above descriptions are merely embodiments of the present application, and are not intended to limit the present application. For those skilled in the art, the present application may have various modifications and variations. Any modification, equivalent replacement, improvement or the like made without departing from the spirit

and principle of the present application should all fall within the scope of claims of the present application.

CLAIMS

1. A word vector processing method, comprising: performing word segmentation on a corpus to obtain words; determining n-gram strokes corresponding to the words, the n-gram stroke
5 representing n successive strokes of a corresponding word; initializing word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words; and after performing the word segmentation, determining the n-gram strokes, and initializing the word vectors and stroke vectors, training the word vectors and the stroke vectors, wherein the training the word vectors and the stroke vectors comprises: determining a designated word in the corpus, and one or more
10 context words of the designated word in the corpus; determining a degree of similarity between the designated word and the context word according to stroke vectors of n-gram strokes corresponding to the designated word as well as a word vector of the context word; selecting one or more words from the words as a negative sample word; determining a degree of similarity between the designated word and each negative sample word; determining a loss characterization
15 value corresponding to the designated word according to a designated loss function, the degree of similarity between the designated word and the context word, and the degree of similarity between the designated word and each negative sample word; and updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value.
20
2. The method of claim 1, wherein determining n-gram strokes corresponding to the words comprises: determining, according to a result of the word segmentation on the corpus, words occurring at least once in the corpus; and performing the following operation on each determined word: determining n-gram strokes corresponding to the word, wherein each n-gram stroke
25 corresponding to the word represents n successive strokes of the word, and n is one positive integer or multiple different positive integers.
3. The method of claim 2, wherein determining, according to a result of the word segmentation on the corpus, words occurring at least once in the corpus specifically comprises: determining,
30 according to the result of the word segmentation on the corpus, a word that occurs in the corpus for not less than a set number of times, the set number of times being not less than 1.

4. The method of claim 1, wherein initializing word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words specifically comprises: initializing the word vectors of the words and the stroke vectors of the n-gram strokes corresponding to the words in a random initialization manner or in a manner of initializing according to a specified probability distribution, wherein stroke vectors of the same n-gram strokes are also the same.
5. The method of claim 1, wherein selecting one or more words from the words as a negative sample word specifically comprises: randomly selecting one or more words from the words as the negative sample word.
6. The method of claim 1, wherein the words are Chinese words, and the word vectors are word vectors of the Chinese words.
7. The method of claim 1, wherein updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value specifically comprises: determining a gradient corresponding to the loss function according to the loss characterization value; and updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the gradient.
8. A non-transitory, computer-readable medium storing one or more instructions executable by a computer system to perform operations comprising: performing word segmentation on a corpus to obtain words; determining n-gram strokes corresponding to the words, the n-gram stroke representing n successive strokes of a corresponding word; initializing word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words; and after performing the word segmentation, determining the n-gram strokes, and initializing the word vectors and stroke vectors, training the word vectors and the stroke vector, wherein the training the word vectors and the stroke vectors comprises: determining a designated word in the corpus, and one or more context words of the designated word in the corpus; determining a degree of similarity between the designated word and the context word according to stroke vectors of n-gram strokes

corresponding to the designated word as well as a word vector of the context word; selecting one or more words from the words as a negative sample word; determining a degree of similarity between the designated word and each negative sample word; determining a loss characterization value corresponding to the designated word according to a designated loss function, the degree
5 of similarity between the designated word and the context word, and the degree of similarity between the designated word and each negative sample word; and updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value.

10 9. The non-transitory, computer-readable medium of claim 8, wherein determining n-gram strokes corresponding to the words comprises: determining, according to a result of the word segmentation on the corpus, words occurring at least once in the corpus; and performing the following operation on each determined word: determining n-gram strokes corresponding to the word, wherein each n-gram stroke corresponding to the word represents n successive strokes of
15 the word, and n is one positive integer or multiple different positive integers.

10. The non-transitory, computer-readable medium of claim 9, wherein determining, according to a result of the word segmentation on the corpus, words occurring at least once in the corpus specifically comprises: determining, according to the result of the word segmentation on the
20 corpus, a word that occurs in the corpus for not less than a set number of times, the set number of times being not less than 1.

11. The non-transitory, computer-readable medium of claim 8, wherein initializing word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words specifically
25 comprises: initializing the word vectors of the words and the stroke vectors of the n-gram strokes corresponding to the words in a random initialization manner or in a manner of initializing according to a specified probability distribution, wherein stroke vectors of the same n-gram strokes are also the same.

30 12. The non-transitory, computer-readable medium of claim 8, wherein determining n-gram strokes corresponding to the words comprises: determining, according to a result of the word

segmentation on the corpus, words occurring at least once in the corpus; and performing the following operation on each determined word: determining n-gram strokes corresponding to the word, wherein each n-gram stroke corresponding to the word represents n successive strokes of the word, and n is one positive integer or multiple different positive integers.

5

13. The non-transitory, computer-readable medium of claim 8, wherein updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value specifically comprises: determining a gradient corresponding to the loss function according to the loss characterization value; and updating the
10 word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the gradient.

14. The non-transitory, computer-readable medium of claim 8, wherein selecting one or more words from the words as a negative sample word specifically comprises: randomly selecting one
15 or more words from the words as the negative sample word.

15. The non-transitory, computer-readable medium of claim 8, wherein the words are Chinese words, and the word vectors are word vectors of the Chinese words.

20 16. A computer-implemented system, comprising: one or more computers; and one or more computer memory devices interoperably coupled with the one or more computers and having tangible, non-transitory, machine-readable media storing one or more instructions that, when executed by the one or more computers, perform one or more operations comprising: performing word segmentation on a corpus to obtain words; determining n-gram strokes corresponding to
25 the words, the n-gram stroke representing n successive strokes of a corresponding word; initializing word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words; and after performing the word segmentation, determining the n-gram strokes, and initializing the word vectors and stroke vectors, training the word vectors and the stroke vector, wherein the training the word vectors and the stroke vectors comprises determining a designated
30 word in the corpus, and one or more context words of the designated word in the corpus; determining a degree of similarity between the designated word and the context word according

to stroke vectors of n-gram strokes corresponding to the designated word as well as a word vector of the context word; selecting one or more words from the words as a negative sample word; determining a degree of similarity between the designated word and each negative sample word; determining a loss characterization value corresponding to the designated word according to a designated loss function, the degree of similarity between the designated word and the context word, and the degree of similarity between the designated word and each negative sample word; and updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value.

17. The computer-implemented system of claim 16, wherein determining, according to a result of the word segmentation on the corpus, words occurring at least once in the corpus specifically comprises: determining, according to the result of the word segmentation on the corpus, a word that occurs in the corpus for not less than a set number of times, the set number of times being not less than 1.

18. The computer-implemented system of claim 17, wherein initializing word vectors of the words and stroke vectors of the n-gram strokes corresponding to the words specifically comprises: initializing the word vectors of the words and the stroke vectors of the n-gram strokes corresponding to the words in a random initialization manner or in a manner of initializing according to a specified probability distribution, wherein stroke vectors of the same n-gram strokes are also the same.

19. The computer-implemented system of claim 16, wherein updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the loss characterization value specifically comprises: determining a gradient corresponding to the loss function according to the loss characterization value; and updating the word vector of the context word and the stroke vectors of the n-gram strokes corresponding to the designated word according to the gradient.

20. The computer-implemented system of claim 16, wherein selecting one or more words from the words as a negative sample word specifically comprises: randomly selecting one or more words from the words as the negative sample word.

- 5 21. The computer-implemented system of claim 16, wherein the words are Chinese words, and the word vectors are word vectors of the Chinese words.