
ENTITY RECOGNIZING METHOD AND APPARATUS OF CHINESE MEDICAL RECORD, DEVICE, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

5 **[01]** The present application declares priority of the Chinese patent application with an application number 201910316061.1, filed on April 19, 2019, entitled “entity recognizing method and apparatus of Chinese medical record, device and storage medium”, entire content of the Chinese patent application is incorporated in the present application by reference.

TECHNICAL FIELD

10 **[02]** The present invention relates to a field of natural language processing, and relates to an entity recognizing method and apparatus of a Chinese medical record, a device and a storage medium.

BACKGROUND

15 **[03]** At present, there is a great demand for the application of named entity recognition in medical records, such as query, search, and sorting of medical records.

20 **[04]** The inventor realizes that the effect of existing Chinese named entity recognition based on deep learning is difficult to improve, and it was previously applied to other languages, such as English. Because of limitations of deep learning models and differences in language characteristics among languages, the application of named entity tasks in Chinese is limited. And because of differences among the general field, other fields and the medical field, application in the direction of medical records in the medical field is limited.

SUMMARY

25 **[05]** The technical problem to be solved by the present invention is to overcome the problem of low accuracy of Chinese named entity recognition based on deep learning in the prior art, an entity recognizing method and apparatus of a Chinese medical record, a device and a storage medium are proposed, corresponding features are extracted from the text content of the Chinese

medical record to convert into feature vectors, then the feature vectors are used as inputs of a model to improve the accuracy of entity recognition rate.

[06] The present invention solves the above technical problems through the following technical solutions:

5 **[07]** An entity recognizing method of a Chinese medical record, including the following steps:

[08] recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same
10 first feature vector;

[09] segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

15 **[10]** recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character;

[11] performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a
20 suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

[12] converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each
25 character corresponding to each character according to a fifth correspondence rule;

[13] splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

[14] inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

[15] The present invention also discloses an entity recognizing apparatus of a Chinese medical record, including:

[16] a first feature vector generating module, recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

[17] a second feature vector generating module, segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

[18] a third feature vector generating module, recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character;

[19] a fourth feature vector generating module, performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

[20] a fifth feature vector generating module, converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule;

[21] a vector set generating module, splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

[22] an entity recognizing module, inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

[23] The present invention also discloses a computing device, which comprises a memory, a processor, and computer programs are stored in the memory, wherein when the computer program is executed by the processor, the following steps are implemented:

[24] recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

[25] segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

[26] recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character;

[27] performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

[28] converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule;

[29] splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

[30] inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

[31] The present invention also discloses a computer-readable storage medium, which stores computer programs, and the computer programs can be executed by at least one processor to implement the following steps:

[32] recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

[33] segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

[34] recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character;

[35] performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

[36] converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule;

[37] splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

[38] inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

[39] In the present invention, entity information in the Chinese medical record is recognized and converted into feature vector firstly, and a vector set converted from the Chinese medical

record is taken as the input of a model, which can effectively improve the accuracy of the model for extracting entities.

BRIEF DESCRIPTION OF THE DRAWINGS

[40] Fig. 1 illustrates a flowchart of an embodiment of an entity recognizing method of a Chinese medical record of the present invention;

[41] Fig. 2 illustrates a structural diagram of an embodiment of an entity recognizing apparatus of a Chinese medical record of the present invention;

[42] Fig. 3 illustrates a diagram of hardware architecture of an embodiment of a computing device of the present invention.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

[43] The present invention will be further described by way of embodiments below, but the present invention is not limited to the scope of the embodiments.

[44] Firstly, the present invention proposes an entity recognizing method of a Chinese medical record.

[45] In one embodiment, as shown in Fig. 1, the entity recognizing method of a Chinese medical record includes the following steps:

[46] Step 01: recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector.

[47] Regular expression matching can be used to identify personal information. The so-called regular expression is a logical formula for string manipulation, which is to use some pre-defined specific characters and a combination of these specific characters to form a "rule string", this "regular string" is used to express a filtering logic for strings.

[48] For example, to identify personal information, it is necessary to create a regular expression for matching personal information firstly, a specific expression depends on the computer programming language used, and each language defines a set of character expression

method; then, the regular expression created is run to match the patient basic information, so that the personal information contained in the patient basic information can be recognized.

[49] The personal information mentioned here mainly refers to patient types and patient ages, the reason for choosing to identify this personal information is that a patient type (male, female, young, old, young, child, infant, etc.) and a patient age can be reflected in each medical record, and corresponding to different patient types and patient ages, doctors may adopt corresponding treatment methods and inspection methods, therefore, identification based on the patient types and the patient ages is conducive to the analysis of a medical record, here, the two characteristics of the patient type and the patient age are selected to be identified from the patient basic information.

[50] Because the patient basic information includes the patient type and the patient age, there are also two first correspondence rules here, namely, a correspondence rule between the patient types and feature vectors and a correspondence rule between the patient ages and feature vectors.

[51] Specifically, there are two types of the correspondence rule between the patient types and the feature vectors: firstly, a length of a feature vector is equal to a number of types of the patient types; each dimension in the feature vector corresponds to one type of the patient types; and the feature vector represents the corresponding patient type through a change of a vector value of a dimension corresponding to the patient type. Secondly, a length of a feature vector is 1; and the feature vector represents different patient types through different vector values.

[52] The following five types of patient types "male, female, elderly, infant and child" are taken as examples to illustrate the two correspondence rules of the patient types and the vector values:

[53] Corresponding to the first correspondence rule: a length of a feature vector is 5, an initial feature vector is assumed to be $[0, 0, 0, 0, 0]$, and each dimension of the feature vector corresponds to a patient type, it is assumed that the patient type corresponding to each dimension in the feature vector from front to back is "male, female, old, infant and child". Based on the above correspondence rules, if the patient type is recognized as "male" in the patient basic information, then the feature value of a first dimension corresponding to the patient type "male" in the initial feature vector changes from 0 to 1, that is, the feature vector $[1,$

0, 0, 0, 0] represents the patient type "male"; if the patient type is recognized as "old", then it is represented by the feature vector [0, 0, 1, 0].

[54] Corresponding to the second correspondence rule: the length of the feature vector is 1, that is, the initial feature vector is [0], different numbers correspond to five patient types, here, it is assumed that the numbers 1, 2, 3, 4 and 5 correspond to the patient types of "male, female, old, infant and child" in turn. Based on the above correspondence rules, if the patient type is recognized as "male" in the patient basic information, then the feature value of the initial feature vector changes from 0 to 1, that is, the feature vector [1] represents the patient type "male", if the patient type is recognized as "old", then it is represented by the feature vector [3].

[55] The correspondence rule of the patient ages and the feature vectors is specifically as follows: the length of the feature vector is 1; the feature vector represents different patient ages through different vector values, and the vector value is equal to a patient age.

[56] The following the patient age as "78 years" is taken as an example to specify the correspondence rule of patient ages and feature vectors. According to the correspondence rule, the length of the feature vector is 1, that is, the initial feature vector is [0], by recognizing the patient age is 78, vector value of the initial feature vector is changed from 0 to 78, that is, the feature vector [78] represents that the patient is 78 years old.

[57] The same first feature vector corresponding to each character in the Chinese medical record refers to: assuming that a patient in a Chinese medical record is recognized as a male, and the age is 78 years old, the feature vector of each word in the Chinese medical record contains the first feature vector [1, 0, 0, 0] or [1] representing the gender as male and the first feature vector [78] representing the age of 78 years old.

[58] Step 02: segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character.

[59] Since it is for Chinese medical records, the word segmentation tools also corresponds to Chinese word segmentation tools, the word segmentation tools mentioned here are all available, and the common ones are jieba, SnowNLP, THULAC, NLPIR, etc., which will not be described in detail.

[60] Single character and words in a sentence are separated by word segmentation, and punctuation can also be separated for subsequent entity recognition.

[61] The word segmentation tools are used to segment the medical records, taking the content of the medical record "直肠腹膜返折上方未及肿块, 结合术前肠镜及病理术中诊断为直肠肛管癌, 决定行 Miles 术" as an example, and "直肠腹膜/返折/上方/未及/肿块/,/结合/术前/肠镜/及/病理/术中诊断/为/直肠肛管癌/,/决定/行/Miles 术/" are obtained after word segmentation.

[62] The second correspondence rule is specifically: a length of a feature vector is 4; the first three dimensions of the feature vector are configured to represent a phrase containing more than two characters, wherein a change of a vector value of a first dimension is configured to represent a character located at the first of the phrase, a change of a vector value of a second dimension is configured to represent a character located in the middle of the phrase, and a change of a vector value of a third dimension is configured to represent a character located at the end of the phrase; a fourth dimension of the feature vector is configured to represent a single-character phrase, and a change of a vector value of the fourth dimension is configured to represent the character in the single-character phrase.

[63] Take the four-character phrase "直肠腹膜" as an example, each character corresponds to an initial feature vector, according to the first correspondence rule, the length of the feature vector is 4, so the initial feature vector corresponding to each word here is [0, 0, 0, 0]. Since the phrase "直肠腹膜" is a four-character phrase, only the first three dimensions of the feature vector are used. "直" is located at the first of the phrase, and the vector value of the first dimension in the initial feature vector is changed accordingly (change from 0 to 1), that is, the feature vector of "直" is [1, 0, 0, 0]; both "肠" and "腹" are located in the middle of the phrase, so the feature vector of these two characters are the same, and vector value of the second dimension in the initial feature vector is changed accordingly (change from 0 to 1), that is, the feature vector of "肠" and "腹" are both [0, 1, 0, 0]; "膜" is located at the end of the phrase, the vector value of the third dimension in the initial feature vector is changed accordingly (change from 0 to 1), that is, the feature vector of "膜" is [0, 0, 1, 0].

[64] Then taking the single-character phrase "及" as an example, the initial feature vector corresponding to the character is also [0, 0, 0, 0], since it is a single-character phrase, only the

fourth dimension of the feature vector is used, that is, vector value of the fourth dimension in the initial feature vector is changed (change from 0 to 1), and the feature vector of "及" is obtained as [0, 0, 0, 1].

[65] Step 03: recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character.

[66] Firstly, each character in the Chinese medical record is matched with a preset radical dictionary, and a matched radical is outputted, so as to identify the radical of each character, wherein the radical dictionary contains association relationships between all Chinese characters and corresponding radicals. Specifically, a radical dictionary containing the association relationships between Chinese characters and corresponding radicals is preset, and each word in the Chinese medical record is matched with the preset radical dictionary, and the matched radical is output.

[67] The third correspondence rule specifically includes two types: firstly, a length of a feature vector is equal to a number of preset entity radicals; each dimension in the feature vector corresponds to one of the entity radicals; the feature vector represents corresponding characters containing the entity radicals through a change of a vector value of a corresponding dimension of the entity radicals; secondly, a length of a feature vector is 1; the feature vector represents characters containing different radicals of the entity correspondingly through different vector values.

[68] The entity radicals are preset according to specific needs, for example, the two most effective entity radicals preset are the sick character box ("疒") and the moon character side ("月"), other radicals can also be added as entity radicals, such as (bamboo prefix "艹", bone character side "骨"), etc.

[69] Take the preset entity radicals, including the sick character box ("疒") and the moon character side ("月") as an example, to specifically explain the two correspondence rules.

[70] Corresponding to the first correspondence rule, a length of a feature vector is 2, and a corresponding initial feature vector is [0, 0], wherein, a change of a vector value of the first dimension is used to represent the sick character box ("疒"), and a change of the vector value of the second dimension is used to represent the moon character side ("month"). Then, for the

last three characters in "直肠腹膜" are the moon character side ("月"), and the moon character side ("月") belongs to the preset entity radicals, so corresponding second feature vectors are all the same [0,1], for the first character "直" in "直肠腹膜" is neither the moon character side ("月") nor the sick character box ("疒"), and it does not belong to the preset entity radicals, then a second feature vector of the character is the initial eigenvector [0, 0]; for the character "病" in "病理" is the sick character box ("疒"), then a corresponding second feature vector is [1, 0].

[71] Corresponding to the second correspondence rule, since a length of a feature vector is fixed to 1, a corresponding initial feature vector is [0], and different vector values 1, 2 are used to represent the sick character box ("疒") and the moon character side ("月") separately. Then, for the last three characters in "直肠腹膜" are the moon character side ("月"), corresponding second feature vectors are all the [2], for the first character "直" in "直肠腹膜" is neither the moon character side ("月") nor the sick character box ("疒"), then a second feature vector of the character is the initial feature vector [0]; and for the character "病" in "病理" is the sick character box ("疒"), then a corresponding second feature vector is [1].

[72] If the preset entity radicals also include the bamboo prefix ("艹") and the bone character side ("骨"), when corresponding to the first correspondence rule, the length of the feature vector is 4, and the initial feature vector is [0, 0,0,0]; and when corresponding to the second correspondence rule, the initial feature vector is [0], and different vector values 1, 2, 3, 4 are used to represent the sick character box ("疒"), the moon character side ("月"), the bamboo prefix ("艹") and the bone character side ("骨") separately, more entity radicals also can be preset, and so on, no more expectation.

[73] Step 04: performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule.

[74] The Chinese medical record need to be preprocessed before traversal, and punctuation match is usually needs to be removed.

[75] n takes a non-zero natural number equals to or less than the length of the Chinese medical record. Preferably, n generally takes the length of the longest phrase in the original

medical dictionary. All of natural numbers less than n are entered to a medical record once. Taking $n=5$ as an example, it is necessary to perform 5-gram traversal, 4-gram traversal, 3-gram traversal, 2-gram traversal, and 1-gram traversal for the Chinese medical record successively.

[76] The so-called n -gram traversal is a common method of natural language processing, in fact, it is word segmentation, and n is the length of each phrase after the word segmentation, the first 5 characters are taken from the first character to form a phrase, the first 5 characters are taken from the second character to form a phrase, and so on, that is, the first character of each phrase is the i -th character in the Chinese medical record, and the last character of each phrase is the $i+n-1$ character in the Chinese medical record, $1 \leq i \leq (\text{the length of the Chinese medical record} - n + 1)$.

[77] Taking the content of the Chinese medical record "直肠腹膜返折上方未及肿块" as an example, the results obtained after 3-gram traversal are "直肠腹", "肠腹膜", "腹膜返", "膜返折", "返折上", "折上方", "上方未", "方未及", "未及肿" and "及肿块".

[78] Any of the existing medical term dictionary may be chosen as the original medical dictionary. In the original medical dictionary, each phrase corresponds to an entity category, and according to entity categories of several phrases separated from the phrase in the prefix dictionary and suffix dictionary follow the entity category corresponding to the phrase in the original medical dictionary.

[79] A construction of the prefix dictionary is specifically: recognizing phrases with more than two characters in the original medical dictionary; storing the first i characters of the recognized phrase into the prefix dictionary in turn, and wherein i is a natural number less than a length of the phrase and greater than half of the length of the phrase, wherein half of the length of the phrase is an integer.

[80] Taking the phrase "左侧粗隆间骨折" in the original medical dictionary as an example, the length of the phrase is 7, half of the length of the phrase is 3.5, rounded to 3, so the value of i is 4, 5, 6, $i=6$ corresponds to "左侧粗隆间骨", $i=5$ corresponds to "左侧粗隆间", $i=4$ corresponds to "左侧粗隆", so the prefix dictionary constructed based on the phrase "左侧粗隆间骨折" includes "左侧粗隆间骨", "左侧粗隆间", and "左侧粗隆".

[81] A construction of the suffix dictionary is specifically: recognizing phrases with more than two characters in the original medical dictionary; storing the last i characters of the

recognized phrase into the suffix dictionary, and wherein i is the natural number less than the length of the phrase and greater than or equal to half of the length of the phrase, wherein half of the length of the phrase is the integer.

[82] Taking the phrase "左侧粗隆间骨折" in the original medical dictionary as an example, the length of the phrase is 7, half of the length of the phrase is 3.5, rounded to 3, so the value of i is 3, 3, 5, 6, $i=6$ corresponds to "侧粗隆间骨折", $i=5$ corresponds to "粗隆间骨折", $i=4$ corresponds to "隆间骨折", and $i=3$ corresponds to "间骨折", so the suffix dictionary constructed based on the phrase "左侧粗隆间骨折" includes "侧粗隆间骨折", "粗隆间骨折", "隆间骨折" and "间骨折".

[83] When matching, the three dictionaries (the original medical dictionary, the prefix dictionary and the suffix dictionary) can be matched at the same time, and a matching order can also be set. No matter which matching method is used, as long as the phrase matches one of the dictionaries, it stops matching with the other two dictionaries. The matching requirement here is an exact match. A matching result includes two types: matching and non-matching. When the matching result is matching, the matching result includes a matching dictionary noun, a matching medical noun, and an entity category corresponding to the medical noun.

[84] The fourth feature vector is used to distinguish the entity categories, and the length of the feature vector is equal to a number of the entity categories, for example, there are six types of entities, representing diseases and diagnoses, symptoms and signs, body parts, examinations and tests, surgery, and drugs, then the length of the corresponding feature vector is 6, then the initial feature vector is $[0, 0, 0, 0, 0, 0]$, and a vector value of each dimension corresponds to an entity category.

[85] When the matching result is non-matching, the initial feature vector is output for each character correspondingly.

[86] When the matching result is matching, according to the entity category of the matched phrase, the vector value of the corresponding dimension of the entity category is changed, according to the matching with the phrase in different dictionaries, change rule is different, therefore, it is necessary to call applicable correspondence rules according to the dictionary nouns contained in the matching result, specifically as follows:

[87] When matching with the original medical dictionary, the applicable correspondence rule is specifically: a length of a feature vector is equal to a number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first, middle or last position of a single character in the phrase by changing an initial vector value to a first vector value, a second vector value or a third vector value.

[88] The following takes the phrase "直肠腹膜" as an example for specific description, wherein, the first vector value, the second vector value and the third vector value are taken as 1, 2, and 3 respectively. The phrase appears in the original medical dictionary and is associated with the entity category of body parts, therefore, these four characters all need to change the vector value of the third dimension. Then according to the specific position of each character in the phrase, "直" is located at the first of the phrase, and the vector value of the third dimension is changed from 0 to 1, that is, the feature vector of "直" is [0, 0, 1, 0, 0, 0]; both "肠" and "腹" are located in the middle of the phrase, so the feature vectors of these two characters are the same, and the vector value of the third dimension vector value is changed from 0 to 2, that is, the feature vectors of "肠" and "腹" are both [0, 0, 2, 0, 0, 0]; "膜" is located at the end of the phrase, and the vector value of the third dimension vector value is changed from 0 to 3. That is, the feature vector of "膜" is [0, 0, 3, 0, 0, 0].

[89] When matching with the prefix dictionary, the applicable correspondence rule is specifically: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first or non-first position of the single character in the phrase by changing the initial vector value to the first vector value or the second vector value.

[90] The following taking the phrase "血细胞" as an example for specific description, wherein, the first vector value and the second vector value are taken as 1, 2 respectively. The phrase appears in the prefix dictionary and is associated with the entity category of examinations and tests, so these three characters all need to change the vector value of the fourth dimension. Then according to the specific position of each word in the phrase, "血" is located at the first place of the phrase, and the vector value of the fourth dimension is changed from 0 to 1, that is, the feature vector of "血" is [0, 0, 0, 1, 0, 0]; both "细" and "胞" are located

in the non-first position of the phrase, and the vector value of the fourth dimension is changed from 0 to 2, that is, the feature vectors of "细" and "胞" are both [0, 0, 0, 2, 0, 0].

[91] When matching with the suffix dictionary, the applicable correspondence rule is specifically: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents to non-last position or the last position of the single character in the phrase by changing the initial vector value to the second vector value or the third vector value.

[92] The following takes the phrase "彩超" as an example for specific description, wherein the second vector value and the third vector value are taken as 2, 3 respectively. The phrase appears in the suffix dictionary and is associated with the entity category of examinations and tests, so these three characters all need to change the vector value of the fourth dimension. Then according to the specific position of each character in the phrase, "彩" is located at the non-last position of the phrase, and the vector value of the fourth dimension is changed from 0 to 2, that is, the feature vector of "彩" is [0, 0, 0, 2, 0, 0]; "超" is located at the end of the phrase, the vector value of the fourth dimension is changed from 0 to 3, that is, the feature vector of "超" is [0, 0, 0, 3, 0, 0].

[93] It should be noted that after n-gram traversal, a sentence will be divided multiple times according to different numbers of characters, therefore, each character will get n feature vectors, but there are only two possibilities for this feature vector, either the original feature vector outputted when non-matching, or the corresponding feature vector outputted when matching (the corresponding feature vector outputted when matching under each traversal is the same). As long as there is a match, the final output corresponding to the character is the corresponding feature vector, unless it does not match each time, the final output corresponding to the character is the initial feature vector.

[94] Step 05: converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule.

[95] The Chinese pinyin conversion tool is an existing technology, and the python package can be used as the conversion tool. The converted pinyin may not indicate tones, and 1, 2, 3,

and 4 may be used to indicate tones, taking "匹" as an example, the converted pinyin can be "pi" or "pi1".

[96] The fifth correspondence rule is specifically: a length of a feature vector is 1; and the feature vector correspondingly represents different pinyin through different vector values.

5 [97] Taking the pinyin "pi" of "匹" as an example, according to the correspondence rule, the length of the feature vector is defined as 1, and the initial value of the feature vector is [0], each pinyin is preset with a corresponding number, assuming that the number corresponding to "pi" is 20, then placing the initial vector value 0 in the feature vector with the 20, that is, the feature vector corresponding to "pi" is [20].

10 [98] Step 06: splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record.

[99] Assuming that the initial vector of a certain character is [0], and the first feature vector, 15 the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector correspond to [1a], [2b], [3c], [4d], [5e] respectively, then the final feature vector of the character is [0, 1a, 2b, 3c, 4d, 5e].

[100] If there are other feature vectors, they can be spliced backwards, and the number is not limited; in addition, the splicing order is not limited to the above.

20 [101] Taking the Chinese medical record containing the following content "性别：男...年龄：78...直肠腹膜..." as an example, the splicing of various feature vectors is further explained.

[102] According to the step 01, there are two first feature vectors corresponding to each character in the Chinese medical record, which represent gender as male [1] (using the second 25 correspondence rule of patient types and feature vectors) and represent age as 78 years old [78] respectively.

[103] According to the step 02, the "直肠腹膜" is a four-character phrase, and the second feature vectors corresponding to these four character s are [1, 0, 0, 0], [0, 1, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0].

[104] According to the step 03, all of the last three characters in the "直肠腹膜" are the moon character side ("月"), so the third feature vectors corresponding to these three characters are [2].

[105] According to step 04, the "直肠腹膜" appears in the original medical dictionary and is associated with the entity category of body parts, therefore, the fourth feature vectors corresponding to these four characters are [0,0,1,0,0,0], [0,0,2,0,0,0], [0,0,2,0,0,0], [0,0,3,0,0,0] respectively.

[106] According to the step 05, the content contained in the Chinese medical record is converted to pinyin, and the fifth feature vector is obtained according to the corresponding number preset for each pinyin, here, it is assumed that the number corresponding to each word in the content "性别: 男.....年龄: 78.....直肠腹膜....." contained in the Chinese medical record is 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, then the fifth feature vector corresponding to each word is [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] respectively.

[107] Assuming that the splicing sequence is from the first feature vector to the fifth feature vector, and the initial vector of each character is [0], then according to the content "性别:

男.....直肠腹膜", the vector set obtained for these characters should be [0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7] [0, 1, 78, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 8] [0, 1, 78, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 9]... ..[0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 14] [0, 1, 78, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0, 0, 15] [0, 1, 78, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0, 0, 16] [0, 1, 78, 0, 0, 1, 0, 2, 0, 0, 3, 0, 0, 0, 17].

[108] Step 07: inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

[109] The model here refers to a Deep Neural Network model, such as two-way LSTM+CRF, and also can be a traditional machine learning model. When the model is trained, input vectors and a corresponding output value are defined for the model, after the model is trained, once an input vector set containing a specific vector value is recognized, the model can recognize the specific entity feature. For example, input the vector set [0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 7] [0, 1, 78, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 8] [0, 1, 78, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 9]... ..[0, 1, 78, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 14] [0, 1, 78, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0, 0, 15] [0, 1, 78, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0, 0, 16] [0, 1, 78, 0, 0, 1, 0, 2, 0, 0, 3, 0, 0, 0, 17], according to the vector values in different dimensions, it can be recognized that the vector set

represents a Chinese medical record of a 78-year-old man, and the last four feature vectors represent a four-character phrase of a body part.

[110] In the embodiment, entities in the Chinese medical record are recognized and converted into feature vectors firstly, and vector set is generated as the input of the model finally, which can effectively improve the accuracy of the model for extracting entities.

[111] Secondly, the present invention provides an entity recognizing apparatus of a Chinese medical record, the device 20 can be divided into one or more modules.

[112] For example, Fig. 2 illustrates a structural diagram of an embodiment of the entity recognizing device 20 of a Chinese medical record. In the embodiment, the device 20 can be divided into a first feature vector generating module 201, a second feature vector generating module 202, a third feature vector generating module 203, a fourth feature vector generating module 204, a fifth feature vector generating module 205, a vector set generating module 206 and an entity recognizing module 207. The following description will specifically introduce the specific functions of the modules 201-207.

[113] The first feature vector generating module 201, recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector.

[114] The second feature vector generating module 202, segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character.

[115] The third feature vector generating module 203, recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character.

[116] The fourth feature vector generating module 204, performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original

medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule.

[117] The fifth feature vector generating module 205, converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule.

[118] The vector set generating module 206, splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record.

[119] The entity recognizing module 207, inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

[120] Thirdly, the present invention also proposes a computing device.

[121] Referring to Fig. 3, it is a schematic diagram of the hardware architecture of an embodiment of a computing device of the present invention. In the embodiment, the computing device 2 is a device capable of automatically performing numerical calculations and/or information processing according to predefined or stored instructions. For example, the computing device 2 may be a smartphone, a tablet, a laptop, a desktop computer, a rack server, a blade server, a tower server, or a cabinet server (including stand-alone servers, or a cluster of multiple servers), and so on. As shown in the figure, the computing device 2 includes, but is not limited to, a memory 21, a processor 22, and a network interface 23 that can be communicated with each other through a system bus. Wherein:

[122] The memory 21 includes at least one type of computer-readable storage medium. The readable storage medium includes a flash memory, a hard disk, a multimedia card, a card type memory (e.g., SD or DX memory, etc.), a random access memory (RAM), a static random access memory (SRAM), a read-only memory (ROM), an electrically erasable programmable read-only memory (EEPROM), a programmable read-only memory (PROM), a magnetic memory, a magnetic disk, an optical disk, etc. In some embodiments, the memory 21 may be

an internal storage unit of the computing device 2 such as a hard disk or a memory of the computing device 2. In other embodiments, the memory 21 may also be an external storage device of the computing device 2, such as a plugged hard disk, a smart media card (SMC), a secure digital (SD) card, a flash memory card, and the like configured on the computing device

2. Of course, the memory 21 may also include both the internal storage unit and the external storage device of the computing device 2. In the embodiment, the memory 21 is generally used to store an operating system and various types of application software installed in the computing device 2 such as computer programs for realizing the entity recognizing method of a Chinese medical record and the like. In addition, the memory 21 may also be used to temporarily store various types of data that have been or will be outputted.

[123] The processor 22, in some embodiments, may be a Central Processing Unit (CPU), a controller, a microcontroller, a microprocessor, or other data processing chip. The processor 22 is generally used to control the overall operation of the computing device 2 such as performing control and processing related to data interaction or communication with the computing device 2. In the embodiment, the processor 22 is used to run program codes stored in the memory 21 or processing data, for example, running computer programs for realizing the entity recognizing method of a Chinese medical record, etc..

[124] The network interface 23 may include a wireless network interface or a wired network interface, the network interface 23 is generally used to establish a communication connection between the computing device 2 and other computing devices. For example, the network interface 23 is used for connecting the computing device 2 to external terminals via a network and establishing a data transmission channel and a communication connection between the computing device 2 and the external terminals. The network can be a wireless or wired network such as an enterprise intranet, the Internet, a Global System of Mobile communication (GSM), a Wideband Code Division Multiple Access (WCDMA), a 4G network, a 5G network, a Bluetooth, Wi-Fi, and the like.

[125] It is to be noted that Fig. 3 only shows a computing device 2 with components 21-23, but it is understood that it is not required to implement all of the shown components and that more or fewer parts can be implemented in lieu thereof.

[126] In the embodiment, the computer programs for realizing the entity recognizing method of a Chinese medical record stored in the memory 21 may be executed by one or more processors (processor 22 in the embodiment) to complete the following steps:

[127] Step 01: recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

[128] Step 02: segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

[129] Step 03: recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character;

[130] Step 04: performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

[131] Step 05: converting each character in the Chinese medical record into pinyin by Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule;

[132] Step 06: splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

[133] Step 07: inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

[134] In addition, the present invention a computer-readable storage medium, the computer-readable storage medium is a non-volatile readable storage medium in which computer programs are stored, and the computer programs can be executed by at least one processor to realize the operation of the entity recognizing method or apparatus of a Chinese medical record.

[135] Wherein, the computer-readable storage medium includes a flash memory, a hard disk, a multimedia card, a card type memory (e.g., SD or DX memory, etc.), a random access memory (RAM), a static random access memory (SRAM), a read-only memory (ROM), an electrically erasable programmable read-only memory (EEPROM), a programmable read-only memory (PROM), a magnetic memory, a magnetic disk, an optical disk, etc. In some embodiments, the computer-readable storage medium may be an internal storage unit of the computing device such as a hard disk or memory of the computing device. In other embodiments, the memory may also be an external storage device of the computing device, such as a plugged hard disk, a smart media card (SMC), a secure digital (SD) card, a flash memory card, and the like configured on the computing device. Of course, the computer-readable storage medium may also include both the internal storage unit and the external storage device of the computing device. In the embodiment, the computer-readable storage medium is generally used to store an operating system and various types of application software installed in the computing device such as computer programs for realizing the entity recognizing method of a Chinese medical record and the like. In addition, the computer-readable storage medium may also be used to temporarily store various types of data that have been or will be outputted.

[136] Although the specific embodiments of the present invention are described above, those skilled in the art should understand that this is only an example, and the protection scope of the present invention is defined by the appended claims. Those skilled in the art can make various changes or modifications to these implementations without departing from the principle and essence of the present invention, but these changes and modifications all fall within the protection scope of the present invention.

WHAT IS CLAIMED IS:

1. An entity recognizing method of a Chinese medical record, comprising the following steps:

recognizing personal information contained in the Chinese medical record, and
5 outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

segmenting the Chinese medical record by a word segmentation tool, taking a phrase
obtained after segmentation as a unit, and outputting a second feature vector for representing a
10 position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

recognizing a radical of each character in the Chinese medical record, and outputting a
third feature vector corresponding to the radical of the each character according to a third
correspondence rule, wherein the third feature vector corresponds to each character;

15 performing n-gram traversal on the Chinese medical record, matching each phrase
obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix
dictionary respectively, and outputting a corresponding fourth feature vector corresponding to
each character according to a matching result and a fourth correspondence rule;

converting each character in the Chinese medical record into pinyin by a Chinese pinyin
20 conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each
character corresponding to each character according to a fifth correspondence rule;

splicing the first feature vector, the second feature vector, the third feature vector, the
fourth feature vector and the fifth feature vector after an initial vector of each character
respectively according to a splicing rule, to obtain a vector set for representing the Chinese
25 medical record; and

inputting the vector set for representing the Chinese medical record into a trained model
to extract entities therein.

2. The entity recognizing method of a Chinese medical record of claim 1, wherein the
30 recognizing personal information contained in the Chinese medical record comprises the
following steps: creating a regular expression for matching the personal information; matching
the regular expression with the patient basic information to identify the personal information
contained in the patient basic information; the first correspondence rule comprises a
correspondence rule between patient types and feature vectors and a correspondence rule

between patient ages and feature vectors; wherein

the correspondence rule between patient types and feature vectors comprises: a length of a feature vector is equal to a number of types of the patient types; each dimension in the feature vector corresponds to one type of the patient types; and the feature vector represents the corresponding patient type through a change of a vector value of a dimension corresponding to the patient type;

or, a length of the feature vector is 1; and the feature vector represents different patient types through different vector values;

the correspondence rule between patient ages and feature vectors comprises: the length of the feature vector is 1; the feature vector represents different patient ages through different vector values, and the vector value is equal to a patient age.

3. The entity recognizing method of a Chinese medical record of claim 1, wherein the second correspondence rule comprises: a length of a feature vector is 4; the first three dimensions of the feature vector are configured to represent a phrase containing more than two characters, wherein a change of a vector value of a first dimension is configured to represent a character located at the first of the phrase, a change of a vector value of a second dimension is configured to represent a character located in the middle of the phrase, and a change of a vector value of a third dimension is configured to represent a character located at the end of the phrase; a fourth dimension of the feature vector is configured to represent a single-character phrase, and a change of a vector value of the fourth dimension is configured to represent the character in the single-character phrase.

4. The entity recognizing method of a Chinese medical record of claim 1, wherein the recognizing a radical of each character in the Chinese medical record, specifically comprises the following steps: matching each character in the Chinese medical record with a preset radical dictionary, and outputting a matched radical, wherein the radical dictionary contains association relationships between all Chinese characters and corresponding radicals;

the third correspondence rule comprises: a length of a feature vector is equal to a number of preset entity radicals; each dimension in the feature vector corresponds to one of the entity radicals; the feature vector represents corresponding characters containing the entity radicals through a change of a vector value of a corresponding dimension of the entity radicals;

or, a length of a feature vector is 1; and the feature vector represents characters containing different radicals of the entity correspondingly through different vector values.

5. The entity recognizing method of a Chinese medical record of claim 1, wherein a construction of the prefix dictionary comprises the following steps: recognizing a phrase with more than two characters in the original medical dictionary; storing the first i characters of the recognized phrase into the prefix dictionary, wherein i is a natural number less than a length of the phrase and greater than half of the length of the phrase, wherein half of the length of the phrase is an integer;

a construction of the suffix dictionary comprises the following steps: recognizing phrases with more than two characters in the original medical dictionary; storing the last i characters of the recognized phrase into the suffix dictionary, wherein i is the natural number less than the length of the phrase and greater than or equal to half of the length of the phrase, wherein half of the length of the phrase is the integer.

6. The entity recognizing method of a Chinese medical record of claim 1, wherein the fourth correspondence rule comprises a correspondence rule applicable when matching with the original medical dictionary, a correspondence rule applicable when matching with the prefix dictionary, and a correspondence rule applicable when matching with the suffix dictionary;

the correspondence rule applicable when matching with the original medical dictionary comprises: a length of a feature vector is equal to a number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first, middle or last position of a single character in the phrase by changing an initial vector value to a first vector value, a second vector value or a third vector value;

the correspondence rule applicable when matching with the prefix dictionary comprises: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first or non-first position of the single character in the phrase by changing the initial vector value to the first vector value or the second vector value;

the correspondence rule applicable when matching with the suffix dictionary comprises: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents to non-last position or the last position of the single character in the phrase by changing the initial vector value to the second vector value or the third vector value.

7. The entity recognizing method of a Chinese medical record of claim 1, wherein the

fifth correspondence rule comprises: a length of a feature vector is 1; and the feature vector correspondingly represents different pinyin through different vector values.

8. An entity recognizing apparatus of a Chinese medical record, comprising:

5 a first feature vector generating module, recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

10 a second feature vector generating module, segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

15 a third feature vector generating module, recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character;

20 a fourth feature vector generating module, performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

25 a fifth feature vector generating module, converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule;

a vector set generating module, splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

30 an entity recognizing module, inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

9. A computing device, which comprises a memory, a processor, and computer programs are stored in the memory, wherein when the computer programs are executed by the processor,

the following steps are implemented:

recognizing personal information contained in the Chinese medical record, and outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

segmenting the Chinese medical record by a word segmentation tool, taking a phrase obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third correspondence rule, wherein the third feature vector corresponds to each character;

performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule;

splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

10. The computing device of claim 9, wherein the recognizing personal information contained in the Chinese medical record comprises the following steps: creating a regular expression for matching the personal information; matching the regular expression with the patient basic information to identify the personal information contained in the patient basic information;

the first correspondence rule comprises a correspondence rule between patient types and feature vectors and a correspondence rule between patient ages and feature vectors;

the correspondence rule between patient types and feature vectors comprises: a length of

a feature vector is equal to a number of types of the patient types; each dimension in the feature vector corresponds to one type of the patient types; and the feature vector represents the corresponding patient type through a change of a vector value of a dimension corresponding to the patient type;

5 or a length of the feature vector is 1; and the feature vector represents different patient types through different vector values;

 the correspondence rule between patient ages and feature vectors comprises: the length of the feature vector is 1; the feature vector represents different patient ages through different vector values, and the vector value is equal to a patient age.

10 11. The computing device of claim 9, wherein the second correspondence rule comprises: a length of a feature vector is 4; the first three dimensions of the feature vector are configured to represent a phrase containing more than two characters, wherein a change of a vector value of a first dimension is configured to represent a character located at the first of the
15 phrase, a change of a vector value of a second dimension is configured to represent a character located in the middle of the phrase, and a change of a vector value of a third dimension is configured to represent a character located at the end of the phrase; a fourth dimension of the feature vector is configured to represent a single-character phrase, and a change of a vector value of the fourth dimension is configured to represent the character in the single-character phrase.

20 12. The computing device of claim 9, wherein the recognizing a radical of each character in the Chinese medical record, specifically comprises the following steps: matching each character in the Chinese medical record with a preset radical dictionary, and outputting a matched radicals, wherein the radical dictionary contains association relationships between all
25 Chinese characters and corresponding radicals;

 the third correspondence rule comprises: a length of a feature vector is equal to a number of preset entity radicals; each dimension in the feature vector corresponds to one of the entity radicals; the feature vector represents corresponding characters containing the entity radicals through a change of a vector value of a corresponding dimension of the entity radicals;

30 or, a length of a feature vector is 1; and the feature vector represents characters containing different radicals of the entity correspondingly through different vector values.

 13. The computing device of claim 9, wherein a construction of the prefix dictionary comprises the following steps: recognizing a phrase with more than two characters in the original

medical dictionary; storing the first i characters of the recognized phrase into the prefix dictionary, wherein i is a natural number less than a length of the phrase and greater than half of the length of the phrase, wherein half of the length of the phrase is an integer;

a construction of the suffix dictionary comprises the following steps: recognizing a phrase with more than two characters in the original medical dictionary; storing the last i characters of the recognized phrase into the suffix dictionary, wherein i is the natural number less than the length of the phrase and greater than or equal to half of the length of the phrase, wherein half of the length of the phrase is the integer.

14. The computing device of claim 9, wherein the fourth correspondence rule comprises a correspondence rule applicable when matching with the original medical dictionary, a correspondence rule applicable when matching with the prefix dictionary, and a correspondence rule applicable when matching with the suffix dictionary;

the correspondence rule applicable when matching with the original medical dictionary comprises: a length of a feature vector is equal to a number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first, middle or last position of a single character in the phrase by changing an initial vector value to a first vector value, a second vector value or a third vector value;

the correspondence rule applicable when matching with the prefix dictionary comprises: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first or non-first position of the single character in the phrase by changing the initial vector value to the first vector value or the second vector value;

the correspondence rule applicable when matching with the suffix dictionary comprises: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents to non-last position or the last position of the single character in the phrase by changing the initial vector value to the second vector value or the third vector value.

15. A non-volatile computer-readable storage medium, which stores computer programs, and the computer programs can be executed by at least one processor to implement the following steps:

recognizing personal information contained in the Chinese medical record, and

outputting a first feature vector corresponding to the personal information according to a first correspondence rule, wherein each word in the Chinese medical record corresponds to the same first feature vector;

segmenting the Chinese medical record by a word segmentation tool, taking a phrase
5 obtained after segmentation as a unit, and outputting a second feature vector for representing a position of each character in the phrase according to a second correspondence rule, wherein the second feature vector corresponds to each character;

recognizing a radical of each character in the Chinese medical record, and outputting a third feature vector corresponding to the radical of the each character according to a third
10 correspondence rule, wherein the third feature vector corresponds to each character;

performing n-gram traversal on the Chinese medical record, matching each phrase obtained after traversal with a preset original medical dictionary, a prefix dictionary, and a suffix dictionary respectively, and outputting a corresponding fourth feature vector corresponding to each character according to a matching result and a fourth correspondence rule;

15 converting each character in the Chinese medical record into pinyin by a Chinese pinyin conversion tool, and outputting a fifth feature vector corresponding to pinyin of the each character corresponding to each character according to a fifth correspondence rule;

splicing the first feature vector, the second feature vector, the third feature vector, the fourth feature vector and the fifth feature vector after an initial vector of each character
20 respectively according to a splicing rule, to obtain a vector set for representing the Chinese medical record;

inputting the vector set for representing the Chinese medical record into a trained model to extract entities therein.

25 16. The computer-readable storage medium of claim 15, wherein the recognizing personal information contained in the Chinese medical record comprises the following steps: creating a regular expression for matching the personal information; matching the regular expression with the patient basic information to identify the personal information contained in the patient basic information; the first correspondence rule comprises a correspondence rule
30 between patient types and feature vectors and a correspondence rule between patient ages and feature vectors; wherein

the correspondence rule between patient types and feature vectors comprises: a length of a feature vector is equal to a number of types of the patient types; each dimension in the feature vector corresponds to one type of the patient types; and the feature vector represents the

corresponding patient type through a change of a vector value of a dimension corresponding to the patient type;

or a length of the feature vector is 1; and the feature vector represents different patient types through different vector values;

5 the correspondence rule between patient ages and feature vectors comprises: the length of the feature vector is 1; the feature vector represents different patient ages through different vector values, and the vector value is equal to a patient age.

10 17. The computer-readable storage medium of claim 15, wherein the second correspondence rule comprises: a length of a feature vector is 4; the first three dimensions of the feature vector are configured to represent a phrase containing more than two characters, wherein a change of a vector value of a first dimension is configured to represent a character located at the first of the phrase, a change of a vector value of a second dimension is configured to represent a character located in the middle of the phrase, and a change of a vector value of a third
15 dimension is configured to represent a character located at the end of the phrase; a fourth dimension of the feature vector is configured to represent a single-character phrase, and a change of a vector value of the fourth dimension is configured to represent the character in the single-character phrase.

20 18. The computer-readable storage medium of claim 15, wherein the recognizing a radical of each character in the Chinese medical record, specifically comprises the following steps: matching each character in the Chinese medical record with a preset radical dictionary, and outputting a matched radical, wherein the radical dictionary contains association relationships between all Chinese characters and corresponding radicals;

25 the third correspondence rule comprises: a length of a feature vector is equal to a number of preset entity radicals; each dimension in the feature vector corresponds to one of the entity radicals; the feature vector represents corresponding characters containing the entity radicals through a change of a vector value of a corresponding dimension of the entity radicals;

30 or, a length of a feature vector is 1; and the feature vector represents characters containing different radicals of the entity correspondingly through different vector values.

19. The computer-readable storage medium of claim 15, wherein a construction of the prefix dictionary comprises the following steps: recognizing a phrase with more than two characters in the original medical dictionary; storing the first *i* characters of the recognized

phrase into the prefix dictionary, wherein i is a natural number less than a length of the phrase and greater than half of the length of the phrase, wherein half of the length of the phrase is an integer;

a construction of the suffix dictionary comprises the following steps: recognizing a phrase with more than two characters in the original medical dictionary; storing the last i characters of the recognized phrase into the suffix dictionary, wherein i is the natural number less than the length of the phrase and greater than or equal to half of the length of the phrase, wherein half of the length of the phrase is the integer.

20. The computer-readable storage medium of claim 15, wherein the fourth correspondence rule comprises a correspondence rule applicable when matching with the original medical dictionary, a correspondence rule applicable when matching with the prefix dictionary, and a correspondence rule applicable when matching with the suffix dictionary;

the correspondence rule applicable when matching with the original medical dictionary comprises: a length of a feature vector is equal to a number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first, middle or last position of a single character in the phrase by changing an initial vector value to a first vector value, a second vector value or a third vector value;

the correspondence rule applicable when matching with the prefix dictionary comprises: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents the first or non-first position of the single character in the phrase by changing the initial vector value to the first vector value or the second vector value;

the correspondence rule applicable when matching with the suffix dictionary comprises: the length of the feature vector is equal to the number of the entity categories; each dimension in the feature vector corresponds to one of the entity categories; the feature vector correspondingly represents to non-last position or the last position of the single character in the phrase by changing the initial vector value to the second vector value or the third vector value.