

METHOD, DEVICE, STORAGE MEDIUM, AND COMPUTING DEVICE FOR SEGMENTING CHINESE WORD BASED ON DEEP LEARNING

[0001] This application claims priority of Chinese Patent Application No. 201910322127.8, entitled “a method, a device, a storage medium, and a computing device for segmenting Chinese word based on deep learning” filed on April 22, 2019 in the China National Intellectual Property Administration (CNIPA), the entire contents of which are incorporated by reference herein.

FIELD

[0002] The present disclosure relates to a technical field of artificial intelligence, specifically a method, a device, a storage medium, and a computing device for segmenting Chinese word based on deep learning.

BACKGROUND

[0003] Current deep learning Chinese word segmentation algorithms are mainly based on cyclic neural network models and their derivative models represented by long short-term memory (LSTM), but the processing ability of the LSTM model in sequence data problems decreases with the increase of the sequence’s length, and there is also a problem of low accuracy of Chinese word segmentation.

SUMMARY

[0004] In view of the above, embodiments of the present disclosure provide a method, a device, a storage medium, and a computing device for segmenting Chinese word based on deep learning to solve the problem of low accuracy of Chinese word segmentation in the prior art.

[0005] In a first aspect, an embodiment of the present disclosure provides a method for segmenting Chinese word based on deep learning. The method includes: converting training corpus data into character-level data; converting the character-level data into sequence data; dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1; extracting a plurality of subsequence data from the i-th data set and inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model; training the i-th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models; converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data; inputting the second data into at least one of the

K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data.

[0006] In a second aspect, an embodiment of the present disclosure provides a device for segmenting Chinese word based on deep learning. The device includes: a first conversion unit for converting training corpus data into character-level data; a second conversion unit for converting the character-level data into sequence data; a first division unit for dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1; a first determining unit for extracting a plurality of subsequence data from the i-th data set and inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model; training the i-th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models; a second determining unit for converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data; inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data.

[0007] In a third aspect, an embodiment of the present disclosure provides a storage medium having programs stored thereon, when the programs are executed by a device containing the storage medium to implement the above-mentioned method for segmenting Chinese word based on deep learning.

[0008] In a fourth aspect, an embodiment of the present disclosure provides a computing device, the computing device includes a memory and a processor, the memory is used to store information including program instructions, the processor is used to control executions of the program instructions, and the program instructions are loaded and executed by the processor to implement the steps of the above-mentioned method for segmenting Chinese word based on deep learning.

[0009] In embodiments of the present disclosure, the target corpus data is converted into character-level data; the character-level data is converted into sequence data; the sequence data is input into the trained temporal convolutional network-conditional random field model, the word segmentation result of the target corpus data is thereby obtained. The temporal convolutional network can expand a receiving field at an exponential growth rate by increasing a number of network layers, so that it can process sequence data with a long sequence length or other data with complex characteristics, and improve an accuracy of a coding result, thereby improving an accuracy of Chinese word segmentation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] In order to describe the technical solutions in embodiments of the present disclosure, the following briefly introduces the accompanying drawings are needed for describing the embodiments. The accompanying drawings in the following description show some embodiments of the present disclosure, and persons of ordinary skill in the art may still derive other drawings from these accompanying drawings without effort.

[0011] FIG. 1 shows a flowchart of a method for segmenting Chinese word based on deep learning provided in an embodiment of the present disclosure.

[0012] FIG. 2 shows a schematic structural diagram of a device for segmenting Chinese word based on deep learning provided in an embodiment of the present disclosure.

[0013] FIG. 3 shows a schematic structural diagram of a computing device provided in an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0014] In order to better understand the technical solutions of the present disclosure, the embodiments of the present disclosure will be described in detail below with reference to the accompanying drawings.

[0015] It should be clear that the described embodiments are only a part of the embodiments of the present disclosure, rather than all of the embodiments. Based on the embodiments in the present disclosure, all other embodiments obtained by a

person of ordinary skill in the art without creative work shall fall within the protection scope of the present disclosure.

[0016] Terms used in the embodiments of the present disclosure are only for the purpose of describing specific embodiments, and are not intended to limit the present disclosure. The singular forms of "a", "said" and "the" used in the embodiments of the present disclosure and the appended claims are also intended to include plural forms, unless the context clearly indicates other meanings.

[0017] It should be understood that the term "and/or" used in the present disclosure is only an association describing associated objects, indicating that there can be three types of relationships. For example, "A and/or B" can mean A alone, A and B at the same time, or B alone. In addition, the forward slash character ("/") in the present disclosure generally indicates that the associated objects can represent themselves and the other.

[0018] FIG. 1 is a flowchart of a method for segmenting Chinese word based on deep learning provided in an embodiment of the present disclosure. As shown in FIG.1, the method includes:

[0019] In block S102, converting training corpus data into character-level data.

[0020] in block S104, converting the character-level data into sequence data.

[0021] In block S106, dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of

subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1. The preset symbols refer to any punctuation mark used for sentence segmentation, such as a period, a question mark, an exclamation mark, a comma, a comma, a semicolon, a colon, etc.

[0022] In block S108, extracting a plurality of subsequence data from the i -th data set and inputting the extracted plurality of subsequence data into the i -th temporal convolutional network-conditional random field model; training the i -th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models.

[0023] In block S110, converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data; inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data.

[0024] The corpus data is a basic resource for learning language knowledge using computer-based materials, and is a language that has actually appeared in the actual use of the language.

[0025] The temporal convolutional network-conditional random field model (TCN-CRF) is a combination model of a temporal convolutional network (TCN) and a conditional random field (CRF). The temporal convolutional network is a

deep learning time convolutional network. The conditional random field is a typical discriminant model. The conditional random field regards word segmentation as a lexeme classification problem, and usually defines lexeme information as: a beginning of a word is “B”, a middle of a word is “M”, an end of a word is “E”, and a single-character word is “S”. A process of word segmentation in the conditional random field is to mark the lexemes, and then segment words between B and E and the word S. For example, if the sentence to be segmented is: "I love Beijing Tiananmen", a result after marking may be: I/S love /S Bei/B Jing/E Tian/B An/M men/E. The result of word segmentation may be: "I/Love/Beijing/Tiananmen.

[0026] In an embodiment of the present disclosure, the target corpus data is converted into character-level data; the character-level data is converted into sequence data; the sequence data is input into the trained temporal convolutional network-conditional random field model, the word segmentation result of the target corpus data is thereby obtained. The temporal convolutional network can expand a receiving field at an exponential growth rate by increasing a number of network layers, so that it can process sequence data with a long sequence length or other data with complex characteristics, and improve an accuracy of a coding result, thereby improving an accuracy of Chinese word segmentation.

[0027] In addition, weightings of neurons on a same feature map in the temporal convolutional network are the same, which can be learned in parallel, and

a processing speed is fast. Therefore, the temporal convolutional network-conditional random field model can also be implemented on a distributed system.

[0028] Optionally, the conversion of the character-level data into sequence data includes: converting the character-level data into sequence data through a preset encoding method, and the preset encoding method is either a One-hot encoding or a word-to-vector encoding.

[0029] The One-hot encoding is also called One-bit effective encoding. The One-hot encoding uses N-bit status registers to encode N states. Each state has an independent register bit, and at any time, only one of them is valid. For example, features of a group of data are in colors, including a yellow, a red, and a green. After using the One-hot encoding, the yellow becomes [100], the red becomes [010], and the green becomes [001], so the sequence data after One-hot encoding corresponds to vectors and can be used in neural network models.

[0030] The word-to-vector encoding can be Word2vec. Word2vec is an efficient algorithm model in which words are represented as real-value vectors. Through training, a processing of text content can be simplified to vector operations in a K-dimensional vector space. Word vectors output by Word2vec can be used for many Neural Linguistic Programming (NLP) related tasks, such as a clustering, finding of synonyms, a part-of-speech analysis, and so on. For example: Word2vec regards character-level data as features, maps the features to the K-dimensional vector space, and obtains sequence data represented by the features.

[0031] Optionally, the input of the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model, and the training of the i-th temporal convolutional network-conditional random field model includes: S1, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model; S2, calculating a loss function value according to the first output data and the extracted plurality of subsequence data; S3, if the loss function value is greater than a preset value, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimizing network parameters of the i-th temporal convolutional network; S4, repeating S1 to S3 until the loss function value is less than or equal to the preset value; S5, if the loss function value is less than or equal to the preset value, terminating the training and obtaining the i-th temporal convolutional network; S6, inputting output data of the trained i-th temporal convolutional network into the i-th conditional random field, training the i-th conditional random field and obtaining the i-th temporal convolutional network-conditional random field model, the i-th conditional random field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

[0032] The training of the i-th temporal convolutional network is based on the loss function value, which specifically includes: initializing network parameters of the i-th temporal convolutional network, using a stochastic gradient descent method to iteratively train the i-th temporal convolutional network, the value of the loss function is calculated once in each iteration, and a loss function value is calculated once in each iteration. Iteration is continued several times until the loss function value reaches a minimum, and the i-th temporal convolutional network and the corresponding converged network parameters are obtained.

[0033] The formula for calculating the loss function value can be:

$$\text{Loss} = \frac{-1}{N} \sum_{i=1}^N [y^{(i)} \log \bar{y}^{(i)} + (1 - y^{(i)}) \log(1 - \bar{y}^{(i)})]$$

[0034] Loss represents the loss function value, N represents a number of subsequence data input to the i-th temporal convolutional network, $y^{(i)}$ represents the i-th subsequence data input to the i-th temporal convolutional network, and $\bar{y}^{(i)}$ represents the data output after inputting the i-th subsequence data into the i-th sequential convolutional neural network.

[0035] Optionally, the training of the i-th conditional random field includes: calculating a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional

network; using a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i -th conditional random field.

[0036] The conditional random field is a Markov random field of a random variable Y under a condition of the given random variable X . A certain random variable of the Markov random field is only related to its neighboring random variables, and has nothing to do with non-adjacent random variables.

[0037] In a conditional probability model $P(Y|X)$, Y is an output variable, which represents a labeled sequence, also called a state sequence, and X is an input variable, which represents an observation sequence that needs to be labeled. A conditional probability model is obtained through a maximum likelihood estimation based on training data. Then the model is used to predict that the output sequence Y will be output when the conditional probability is maximum, for a given input sequence X . A linear chain conditional random field is a commonly used conditional random field, an input sequence $X=(X_1, X_2, \dots, X_n)$, an output sequence $Y=(Y_1, Y_2, \dots, Y_n)$, which are the random variable sequence represented by the linear chain. If the random variable sequence X is given, the conditional probability distribution $P(Y|X)$ of the random variable sequence Y constitutes a conditional random field.

[0038] Maximum likelihood estimation refers to obtaining experimental results through several experiments, and obtaining a certain parameter value according to the experimental results to maximize a probability of sample appearances.

Maximum likelihood estimation provides a method for evaluating model parameters given observation data, that is, "the model has been determined and the parameters are unknown". Given that the sample data is $X=(X_1, X_2, \dots, X_n)$, n is a number of the sample data, an estimated parameter t , and a likelihood function of t relative to X is $f(t)=P(X|t)=P(X_1, X_2, \dots, X_n|t)=\prod_{i=1}^n P(X_i|t)$, wherein i is a natural number from 1 to n . If t' is a value of t that can maximize the likelihood function $f(t)$ in a parameter space, then t' should be a "most likely" parameter, and t' is a maximum likelihood estimator of t .

[0039] Optionally, the inputting of the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data, includes: dividing the second data according to the preset symbols to obtain a plurality of sequence data; grouping the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$; according to a length of the subsequence data used in the training process, selecting L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L_1 -st to L_L -th temporal convolutional network-conditional random field models; inputting all the sequence data contained in the j -th data set into the L_j -th trained temporal

convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the L_j -th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j -th data set, j taking a natural number between 1 and L in turn, and L_j being a natural number between 1 and K ; combining the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.

[0040] For example, suppose the value of K is 5, and the lengths of subsequences used in the training of the 5 temporal convolutional network-conditional random field models are 10, 20, 30, 40, and 50, respectively. After dividing the second data, two sequence data are obtained. The lengths of the two sequence data are 20 and 50. According to the lengths of 20 and 50 of the subsequence data used in the training process, two trained temporal convolutional network-conditional random field models are selected from the five trained temporal convolutional network-conditional random field models. The length of the subsequence data used in the training process of the first selected trained temporal convolutional network-conditional random field model is 20. The length of the subsequence data used in the training process of the second selected trained temporal convolutional network-conditional random field model is 50. The sequence data with the length of 20 is input into the first trained temporal convolutional network-conditional random field model, and word segmentation

results are obtained. The sequence data with the length of 50 is input into the second trained temporal convolutional network-conditional random field model, and word segmentation results are obtained. The word segmentation results output by the first trained temporal convolutional network-conditional random field model and the word segmentation results output by the second trained temporal convolutional network-conditional random field model are then combined to obtain the word segmentation result of the target corpus data.

[0041] FIG. 2 is a schematic structural diagram of a device for segmenting Chinese words based on deep learning provided in one embodiment of the present disclosure. The device is used to execute the above-mentioned method device for segmenting Chinese word based on deep learning. As shown in FIG. 2, the device includes: a first conversion unit 10, a second conversion unit 20, a first division unit 30, a first determination unit 40, and a second determination unit 50.

[0042] The first conversion unit 10, is configured to convert training corpus data into character-level data.

[0043] The second conversion unit 20, is configured to convert the character-level data into sequence data.

[0044] The first division unit 30, is configured to divide the sequence data according to preset symbols to obtain a plurality of subsequence data; group the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being

the same, and K being a natural number greater than 1. The preset symbols refers to any punctuation mark used for sentence segmentation, such as a period, a question mark, an exclamation mark, a comma, a comma, a semicolon, a colon, etc.

[0045] The first determination unit 40 is configured to extract a plurality of subsequence data from the i -th data set and input the extracted plurality of subsequence data into the i -th temporal convolutional network-conditional random field model; train the i -th temporal convolutional network-conditional random field model, and take a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models.

[0046] The second determination unit 50 is configured to convert target corpus data into character-level data as being first data, and convert the first data into sequence data as being second data; input the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtain a word segmentation result of the target corpus data.

[0047] The corpus data is a basic resource for learning language knowledge using computer-based materials, and is a language that has actually appeared in the actual use of the language.

[0048] The temporal convolutional network-conditional random field model (TCN-CRF) is a combination model of a temporal convolutional network (TCN) and a conditional random field (CRF). The temporal convolutional network is a deep learning time convolutional network. The conditional random field is a typical

discriminant model. The conditional random field regards word segmentation as a lexeme classification problem, and usually defines lexeme information as: a beginning of a word is “B”, a middle of a word is “M”, an end of a word is “E”, and a single-character word is “S”. A process of word segmentation in the conditional random field is to mark the lexemes, and then segment words between B and E and the word S. For example, if the sentence to be segmented is: "I love Beijing Tiananmen", a result after marking may be: I/S love /S Bei/B Jing/E Tian/B An/M men/E. The result of word segmentation may be: "I/Love/Beijing/Tiananmen.

[0049] In an embodiment of the present disclosure, the target corpus data is converted into character-level data; the character-level data is converted into sequence data; the sequence data is input into the trained temporal convolutional network-conditional random field model, the word segmentation result of the target corpus data is thereby obtained. The temporal convolutional network can expand a receiving field at an exponential growth rate by increasing a number of network layers, so that it can process sequence data with a long sequence length or other data with complex characteristics, and improve an accuracy of a coding result, thereby improving an accuracy of Chinese word segmentation.

[0050] In addition, weightings of neurons on a same feature map in the temporal convolutional network are the same, which can be learned in parallel, and a processing speed is fast. Therefore, the temporal convolutional network-conditional random field model can also be implemented on a distributed system.

[0051] Optionally, the second determination unit 20 includes a conversion subunit. The conversion subunit is used to convert the character-level data into sequence data through a preset encoding method, and the preset encoding method is either a One-hot encoding or a word-to-vector encoding.

[0052] Optionally, the first determination unit 40 is configured to perform the following steps: S1, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model; S2, calculating a loss function value according to the first output data and the extracted plurality of subsequence data; S3, if the loss function value is greater than a preset value, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimizing network parameters of the i-th temporal convolutional network; S4, repeating S1 to S3 until the loss function value is less than or equal to the preset value; S5, if the loss function value is less than or equal to the preset value, terminating the training and obtaining the i-th temporal convolutional network; S6, inputting output data of the trained i-th temporal convolutional network into the i-th conditional random field, training the i-th conditional random field and obtaining the i-th temporal convolutional network-conditional random field model, the i-th conditional random

field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

[0053] Optionally, the first determination unit 40 includes a first calculating subunit and a first determining subunit. The first calculating subunit is used to calculate a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional network. The first determining subunit uses a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i-th conditional random field.

[0054] Optionally, the second determining unit 50 includes a division subunit, a grouping subunit, a second determining subunit, and a combining subunit. The division subunit divides the second data according to the preset symbols to obtain a plurality of sequence data. The grouping subunit groups the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$. According to a length of the subsequence data used in the training process, the second determining subunit selects L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L1-st to LL-th temporal convolutional network-conditional random field models; inputs all the sequence data contained in the j-th data set into the Lj-th trained temporal

convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the L_j -th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j -th data set, j taking a natural number between 1 and L in turn, and L_j being a natural number between 1 and K . The combining subunit combines the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.

[0055] One embodiment of the present disclosure provides a storage medium having programs stored thereon, when the programs are executed by a device containing the storage medium to implement the following steps: converting training corpus data into character-level data; converting the character-level data into sequence data; dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1; extracting a plurality of subsequence data from the i -th data set and inputting the extracted plurality of subsequence data into the i -th temporal convolutional network-conditional random field model; training the i -th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models; converting target corpus

data into character-level data as being first data, and converting the first data into sequence data as being second data; inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data.

[0056] Optionally, when the programs are executed by the device containing the storage medium, implementing the following steps: converting the character-level data into sequence data through a preset encoding method, and the preset encoding method is either a One-hot encoding or a word-to-vector encoding.

[0057] Optionally, when the programs are executed by the device containing the storage medium, also implementing the following steps: S1, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model; S2, calculating a loss function value according to the first output data and the extracted plurality of subsequence data; S3, if the loss function value is greater than a preset value, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimizing network parameters of the i-th temporal convolutional network; S4, repeating S1 to S3 until the loss function value is less than or equal to the preset value; S5, if the loss function value is less than or equal to the preset value, terminating the training and obtaining the i-th temporal convolutional

network; S6, inputting output data of the trained i-th temporal convolutional network into the i-th conditional random field, training the i-th conditional random field and obtaining the i-th temporal convolutional network-conditional random field model, the i-th conditional random field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

[0058] Optionally, when the programs are executed by the device containing the storage medium, also implementing the following steps: calculating a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional network; using a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i-th conditional random field.

[0059] Optionally, when the programs are executed by the device containing the storage medium, also implementing the following steps: dividing the second data according to the preset symbols to obtain a plurality of sequence data; grouping the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$; according to a length of the subsequence data used in the training process, selecting L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L1-st to LL-th temporal convolutional network-conditional random field models; inputting all the

sequence data contained in the j -th data set into the L_j -th trained temporal convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the L_j -th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j -th data set, j taking a natural number between 1 and L in turn, and L_j being a natural number between 1 and K ; combining the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.

[0060] One embodiment of the present disclosure provides a computing device, the computing device including a storage device and a processor, and the processor executes programs stored in the storage device to implement the following steps: converting training corpus data into character-level data; converting the character-level data into sequence data; dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1; extracting a plurality of subsequence data from the i -th data set and inputting the extracted plurality of subsequence data into the i -th temporal convolutional network-conditional random field model; training the i -th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal

convolutional network-conditional random field models; converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data; inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data.

[0061] Optionally, the processor executes programs stored in the storage device to also convert the character-level data into sequence data through a preset encoding method, the preset encoding method being either a One-hot encoding or a word-to-vector encoding.

[0062] Optionally, the processor executes programs stored in the storage device to also implement the steps of S1, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model; S2, calculating a loss function value according to the first output data and the inputted the extracted plurality of subsequence data; S3, if the loss function value is greater than a preset value, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimizing network parameters of the i-th temporal convolutional network; S4, repeating S1 to S3 until the loss function value is less than or equal to the preset value; S5, if the loss function value is less than or equal to the preset

value, terminating the training and obtaining the i-th temporal convolutional network; S6, inputting output data of the trained i-th temporal convolutional network into the i-th conditional random field, training the i-th conditional random field and obtaining the i-th temporal convolutional network-conditional random field model, the i-th conditional random field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

[0063] Optionally, the processor executes programs stored in the storage device to also implement the steps of calculating a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional network and using a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i-th conditional random field.

[0064] Optionally, the processor executes programs stored in the storage device to also implement the following steps: dividing the second data according to the preset symbols to obtain a plurality of sequence data; grouping the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$; according to a length of the subsequence data used in the training process, selecting L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L1-st to LL-th temporal convolutional

network-conditional random field model; inputting all the sequence data contained in the j -th data set into the L_j -th trained temporal convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the L_j -th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j -th data set, j taking a natural number between 1 and L in turn, and L_j being a natural number between 1 and K ; and combining the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.

[0065] FIG. 3 is a schematic structural diagram of a computing device provided in an embodiment of the present disclosure. The computing device 50 of the embodiment includes a processor 51, a memory 52, and computer programs 53 stored in the memory 52 and running on the processor 51. The computer programs 53 are executed by the processor 51 to implement the method for segmenting Chinese word based on deep learning in the embodiments. In order to avoid repetition, it will not be repeated here. Or, when the computer programs are executed by the processor 51 to implement the function of each model/unit in the device for segmenting Chinese word based on deep learning in the embodiments. In order to avoid repetition, it will not be repeated here.

[0066] The computing device 50 can be a computing device such as a desktop computer, a notebook, a palmtop computer, and a cloud server. The computing

device 50 can include, but is not limited to, a processor 51 and a memory 52. Those skilled in the art will understand that the schematic diagram 3 is only an example of the computing device 50, and does not constitute a limitation on the computing device 50. Another computing device 50 may include more or fewer components than shown in the figures, or combine some components or have different components. For example, the computing device 50 may further include an input/output device, a network access device, a bus, and the like.

[0067] The processor 51 can be a central processing unit (CPU), or can be other general-purpose processor, digital signal processor (DSPs), and application specific integrated circuit (ASIC), Field-Programmable Gate Array (FPGA), or other programmable logic device, discrete gate, or transistor logic device, or discrete hardware component, etc.. The general-purpose processor may be a microprocessor or the processor may also be any conventional processor or the like.

[0068] The memory 52 can be an internal storage unit of the computer device 50, such as a hard disk or a memory of the computer device 50. The memory 52 can also be an external storage device of the computer device 50, such as a plug-in hard disk, a smart media card (SMC), and a secure digital (SD) Card, a flash card equipped on the computer device 50. Further, the memory 52 may also include both an internal storage unit of the computer device 50 and an external storage device. The memory 52 is used to store computer programs and other programs and data

required by the computer equipment. The memory 52 can also be used to temporarily store data that has been output or will be output.

[0069] Those skilled in the art can clearly understand that, for the convenience and conciseness of the description, the specific working processes of the above-described system, device, and computing unit can refer to the corresponding process in the foregoing method embodiment.

[0070] In the several embodiments provided in the present disclosure, it should be understood that the disclosed system, device, and method may be implemented in other ways. For example, the device embodiments described above are merely illustrative. For example, the division of the units is only a logical function division, and there may be other divisions in actual implementation. For example, a plurality of units or components can be combined or can be integrated into another system. Some features can be ignored or not implemented. In addition, a displayed or discussed mutual coupling or direct coupling or communication connection can mean indirect coupling or communication connection through interfaces, devices, or units, and can be in electrical, mechanical, or other form.

[0071] The above descriptions are only the preferred embodiments of the present disclosure and are not intended to limit the present disclosure. Any modification, equivalent replacement, improvement, etc. made within the spirit and principle of the present disclosure shall be included within the scope of protection of the present disclosure.

CLAIMS

We claim:

1. A method for segmenting Chinese word based on deep learning, the method comprising:

converting training corpus data into character-level data;

converting the character-level data into sequence data;

dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1;

extracting a plurality of subsequence data from the i-th data set and inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model;

training the i-th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models;

converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data;

inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models; and

obtaining a word segmentation result of the target corpus data.

2. The method of claim 1, wherein the converting the character-level data into sequence data comprises:

converting the character-level data into sequence data through a preset encoding method, the preset encoding method is either a One-hot encoding or a word-to-vector encoding.

3. The method of claim 1, wherein the inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model, and training the i-th temporal convolutional network-conditional random field model comprises:

S1, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model;

S2, calculating a loss function value according to the first output data and the extracted plurality of subsequence data;

S3, if the loss function value is greater than a preset value, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimizing network parameters of the i-th temporal convolutional network;

S4, repeating S1 to S3 until the loss function value is less than or equal to the preset value;

S5, if the loss function value is less than or equal to the preset value, terminating the training and obtaining the i-th temporal convolutional network; and

S6, inputting output data of the trained i-th temporal convolutional network into the i-th conditional random field, training the i-th conditional random field and obtaining the i-th temporal convolutional network-conditional random field model, the i-th conditional random field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

4. The method of claim 3, wherein the training the i-th conditional random field comprises:

calculating a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional network; and

using a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i-th conditional random field.

5. The method of any claim 1-4, wherein the inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data, comprises:

dividing the second data according to the preset symbols to obtain a plurality of sequence data;

grouping the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$;

according to a length of the subsequence data used in the training process, selecting L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L1-st to LL-th temporal convolutional network-conditional random field models; inputting all the sequence data contained in the j-th data set into the Lj-th trained temporal convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the Lj-th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j-th data set, j taking a natural number between 1 and L in turn, and Lj being a natural number between 1 and K; and

combining the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.

6. A device for segmenting Chinese word based on deep learning, the device comprising:

a first conversion unit, configured to convert training corpus data into character-level data;

a second conversion unit, configured to convert the character-level data into sequence data;

a first division unit, configured to divide the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1;

a first determination unit, configured to extract a plurality of subsequence data from the i-th data set and inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model; training the i-th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models; and

a second determination unit, configured to converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data; inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data.

7. The device of claim 6, wherein the second conversion unit comprises:

a conversion subunit configured to convert the character-level data into sequence data through a preset encoding method, the preset encoding method is either a One-hot encoding or a word-to-vector encoding

8. The device of claim 6, wherein the first determining unit is further configured to:

S1, input the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model;

S2, calculate a loss function value according to the first output data and the extracted plurality of subsequence data;

S3, if the loss function value is greater than a preset value, input the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimize network parameters of the i-th temporal convolutional network;

S4, repeat S1 to S3 until the loss function value is less than or equal to the preset value;

S5, if the loss function value is less than or equal to the preset value, terminate the training and obtaining the i-th temporal convolutional network; and

S6, input output data of the trained i-th temporal convolutional network into the i-th conditional random field, train the i-th conditional random field and obtain the i-th temporal convolutional network-conditional random field model, the i-th conditional random field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

9. The device of claim 8, wherein the first determination unit comprises:

a first calculation subunit configured to calculate a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional network; and

a first determination subunit configured to use a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i-th conditional random field.

10. The device of any claim 6-9, wherein the second determination unit comprises:

a division subunit configured to divide the second data according to the preset symbols to obtain a plurality of sequence data;

a grouping subunit configured to group the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$;

a second determining subunit configured to, according to a length of the subsequence data used in the training process, select L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L1-st to LL-th temporal convolutional network-conditional random field models; input all the sequence data contained in the j-th data set into the Lj-th trained temporal convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the Lj-th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j-th data set, j taking a natural number between 1 and L in turn, and Lj being a natural number between 1 and K; and

a combining subunit configured to combine the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.

11. A storage medium having programs stored thereon, when the programs are executed by a device containing the storage medium to implement the following steps:

converting training corpus data into character-level data;

converting the character-level data into sequence data;

dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the

subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1;

extracting a plurality of subsequence data from the i-th data set and inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model; training the i-th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models;

converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data;

inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models; and

obtaining a word segmentation result of the target corpus data.

12. The storage medium of claim 11, wherein when the programs are executed by the device containing the storage medium to implement the step of converting the character-level data into sequence data, comprises:

converting the character-level data into sequence data through a preset encoding method, the preset encoding method is either a One-hot encoding or a word-to-vector encoding

13. The storage medium of claim 11, wherein when the programs are executed by the device containing the storage medium to implement the step of inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model, and training the i-th temporal convolutional network-conditional random field model, comprises:

S1, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model;

S2, calculating a loss function value according to the first output data and the extracted plurality of subsequence data;

S3, if the loss function value is greater than a preset value, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimizing network parameters of the i-th temporal convolutional network;

S4, repeating S1 to S3 until the loss function value is less than or equal to the preset value;

S5, if the loss function value is less than or equal to the preset value, terminating the training and obtaining the i-th temporal convolutional network; and

S6, inputting output data of the trained i-th temporal convolutional network into the i-th conditional random field, training the i-th conditional random field and obtaining the

i-th temporal convolutional network-conditional random field model, the i-th conditional random field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

14. The storage medium of claim 13, wherein when the programs are executed by the device containing the storage medium to implement the step of training the i-th conditional random field, comprises:

calculating a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional network; and

using a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i-th conditional random field.

15. The storage medium of any claim 11-14, wherein when the programs are executed by the device containing the storage medium to implement the step of inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target corpus data, comprises:

dividing the second data according to the preset symbols to obtain a plurality of sequence data;

grouping the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$;

according to a length of the subsequence data used in the training process, selecting L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L1-st to LL-th temporal convolutional network-conditional random field models; inputting all the sequence data contained in the j-th data set into the Lj-th trained temporal convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the Lj-th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j-th data set, j taking a natural number between 1 and L in turn, and Lj being a natural number between 1 and K; and

combining the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.

16. A computing device comprising a memory and a processor, the memory is used to store information including program instructions, the processor is used to control executions of the program instructions, and the program instructions are loaded and executed by the processor to implement the following steps:

converting training corpus data into character-level data;

converting the character-level data into sequence data;

dividing the sequence data according to preset symbols to obtain a plurality of subsequence data; grouping the plurality of subsequence data according to lengths of the subsequence data to obtain K data sets, a length of subsequence data contained in each of the K data sets being the same, and K being a natural number greater than 1;

extracting a plurality of subsequence data from the i-th data set and inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model;

training the i-th temporal convolutional network-conditional random field model, and taking a natural number between 1 and K in sequence for i to obtain K temporal convolutional network-conditional random field models;

converting target corpus data into character-level data as being first data, and converting the first data into sequence data as being second data;

inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models; and

obtaining a word segmentation result of the target corpus data.

17. The computing device of claim 16, wherein the step of converting the character-level data into sequence data when the program instructions are loaded and executed by the processor comprises:

converting the character-level data into sequence data through a preset encoding method, the preset encoding method is either a One-hot encoding or a word-to-vector encoding

18. The computing device of claim 16, wherein the step of inputting the extracted plurality of subsequence data into the i-th temporal convolutional network-conditional random field model, and training the i-th temporal convolutional network-conditional random field model when the program instructions are loaded and executed by the processor comprises:

S1, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for forward propagation to obtain a first output data, the i-th temporal convolutional network being a temporal convolutional network in the i-th temporal convolutional network-conditional random field model;

S2, calculating a loss function value according to the first output data and the extracted plurality of subsequence data;

S3, if the loss function value is greater than a preset value, inputting the extracted plurality of subsequence data into the i-th temporal convolutional network for back propagation, and optimizing network parameters of the i-th temporal convolutional network;

S4, repeating S1 to S3 until the loss function value is less than or equal to the preset value;

S5, if the loss function value is less than or equal to the preset value, terminating the training and obtaining the i-th temporal convolutional network; and

S6, inputting output data of the trained i-th temporal convolutional network into the i-th conditional random field, training the i-th conditional random field and obtaining the i-th temporal convolutional network-conditional random field model, the i-th conditional random field being a conditional random field in the i-th temporal convolutional network-conditional random field model.

19. The computing device of claim 18, wherein the step of training the i-th conditional random field when the program instructions are loaded and executed by the processor comprises:

calculating a conditional probability of the output data of the i-th conditional random field according to the data output by the trained i-th temporal convolutional network; and

using a maximum likelihood estimation method to train a maximum value of the conditional probability of the output data of the i-th conditional random field.

20. The computing device of any claim 16-19, wherein the step of inputting the second data into at least one of the K trained temporal convolutional network-conditional random field models, and obtaining a word segmentation result of the target

corpus data when the program instructions are loaded and executed by the processor, comprises:

dividing the second data according to the preset symbols to obtain a plurality of sequence data;

grouping the plurality of sequence data according to lengths of the sequence data to obtain L data sets, a length of sequence data contained in each of the L data sets being the same, and L being a natural number, $1 \leq L \leq K$;

according to a length of the subsequence data used in the training process, selecting L trained temporal convolutional network-conditional random field models from K trained temporal convolutional network-conditional random field models, to obtain L1-st to LL-th temporal convolutional network-conditional random field models; inputting all the sequence data contained in the j-th data set into the Lj-th trained temporal convolutional network-conditional random field model to obtain a plurality of word segmentation results, a length of the subsequence data used in the training process of the Lj-th temporal convolutional network-conditional random field model being equal to a length of the sequence data contained in the j-th data set, j taking a natural number between 1 and L in turn, and Lj being a natural number between 1 and K; and

combining the plurality of word segmentation results to obtain the word segmentation result of the target corpus data.