Description

Title of The Inventions: Text Segmentation Method, Text Segmentation Apparatus, Electronic Device, and Computer-Readable Storage Medium

Cross Reference to Related Applications

The present disclosure claims priority to Chinese Patent Application No. 202510210103.9, filed with the Chinese Patent Office on February 25, 2025, entitled “Text Segmentation Method, Text Segmentation Apparatus, Electronic Device, and Computer-readable Storage Medium”, the entirety of which is incorporated herein by reference.

Technical Field

The present disclosure relates to the field of computer technology, and in particular to a text segmentation method, a text segmentation apparatus, an electronic device, and a computer-readable storage medium.

Background Art

With the breakthrough progress of large models in the field of natural language processing, question answering systems based on large models have overcome the defects of traditional question answering systems that cannot effectively deal with diverse questions from users and make mechanical responses, because they can understand more complex language structures and capture more abundant contextual information, so as to provide users with more accurate, fluent, and personalized answers and interact with users in a more humanized and intelligent manner.

Although the response quality of question answering systems is improved by large models, there is a defect that the generated answers, which although seem reasonable and logical, are actually inaccurate or contrary to the facts, due to the limited coverage of specific domain knowledge in their training processes. Therefore, based on a Retrieval-Augmented Generation (simply referred to as RAG) method, question answering systems based on large models are provided with strong support by combining external knowledge retrieval with generative large models. The RAG method enhances the knowledge coverage range of a large model and improves its accuracy and reliability by firstly retrieving relevant information from a database and inputting the retrieved knowledge into the model as context before generating an answer.

The information in the database usually comes from an external document. It is necessary to firstly segment the external document into a plurality of fragments, and then convert each fragment into a vector and store them in the database. The reasonableness of the segmentation of the external document will affect the integrity of the information in the context retrieved by the RAG, and thus affect the accuracy and reliability of a question-answering output fed back by the RAG.

Summary

One of the objects of the present disclosure is to provide a text segmentation method, a text segmentation apparatus, an electronic device, and a computer-readable storage medium, which can reasonably segment a text semantically, ensure the semantic coherence of the segmented semantic segments, and thus improve the accuracy and reliability of a question-answering output fed back by RAG.

In order to achieve the above object, the following technical solutions are employed in embodiments of the present disclosure.

In a first aspect, the present disclosure provides a text segmentation method, the method comprising:

acquiring a plurality of sentences obtained after sentence-level segmentation of a target text;

sliding a preset sliding window multiple times by using the first sentence of the plurality of sentences as a starting point and using the last sentence of the plurality of sentences as an end point, and semantically segmenting a plurality of target sentences located in the preset sliding window after each sliding operation to obtain at least one semantic segment; and

clustering, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained.

In an optional implementation mode, the step of semantically segmenting a plurality of target sentences located in the preset sliding window to obtain at least one semantic segment comprises:

inputting the plurality of target sentences into a pretrained text segmentation model to obtain semantic continuity of sentence pairs composed of every two adjacent target sentences;

affiliating, for any target sentence pair, the two target sentences in the target sentence pair to the same semantic segment if the semantic continuity of the target sentence pair is greater than a preset value, and if not, affiliating the two target sentences in the target sentence pair to different semantic segments; and

sequentially treating each sentence pair as the target sentence pair to obtain at least one semantic segment after semantic segmentation of the plurality of target sentences.

In an optional implementation mode, the first target sentence in the preset sliding window after the current sliding operation is the same as the last target sentence in the preset sliding window in the previous sliding operation, and every two adjacent target sentences in the preset sliding window form a sentence pair;

the step of clustering, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments

obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs comprises:

acquiring semantic continuity of a first sentence pair in the preset sliding window after the current sliding operation;

calculating semantic similarity between the two target sentences in the first sentence pair;

calculating a semantic association score of the first sentence pair according to the semantic continuity and the semantic similarity between the two target sentences in the first sentence pair;

merging the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation into one semantic segment if the semantic association score is greater than a preset score, and if not, treating the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation as different semantic segments.

In an optional implementation mode, the calculation of the semantic similarity between the two target sentences in the first sentence pair comprises: extracting feature vectors of the two target sentences, respectively; calculating cosine similarity between the feature vectors of the two target sentences, and using the calculated cosine similarity as the semantic similarity.

In an optional implementation mode, the calculation of the semantic similarity between the two target sentences in the first sentence pair comprises: inputting the two target sentences into a multimodal multi-task embedding model to obtain the semantic similarity between the two target sentences.

In an optional implementation mode, the semantic association score of the first sentence pair is calculated by the following formula: semantic association score = semantic continuity × continuity weight + semantic similarity × similarity weight.

In an optional implementation mode, the step of semantically segmenting a plurality of target sentences located in the preset sliding window to obtain at least one semantic segment comprises: acquiring semantic continuity of a sentence pair composed of every two adjacent target sentences; calculating semantic similarity of the sentence pair composed of every two adjacent target sentences; calculating a semantic association score of the sentence pair composed of every two adjacent target sentences according to the semantic continuity and the semantic similarity between the two target sentences in the sentence pair composed of every two adjacent target sentences; if the semantic association score is greater than a preset score, affiliating the target sentences in the corresponding sentence pair to the same semantic segment, and if not, affiliating them to different semantic segments.

In an optional implementation mode, after the step of clustering, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset

sliding window after the previous sliding operation and after the current sliding operation belongs, the method further comprises:

acquiring a preset restrictive condition, where the preset restrictive condition is configured to restrict semantic segments to be inputted into a large model in a retrieval-augmented generation (RAG) system;

processing each of the target semantic segments according to the preset restrictive condition to obtain a reference semantic segment corresponding to each of the target semantic segments, and storing each of the reference semantic segments into a vector database in a vectorized form, so that when the large model receives input data, a semantic segment vector related to the input data is retrieved from the vector database, and feedback is performed based on the input data and the retrieved semantic segment vector.

In an optional implementation mode, the preset restrictive condition includes a maximum length of a single semantic segment to be inputted into a large model and/or a degree of overlap between two adjacent semantic segments to be inputted into the large model.

In an optional implementation mode, the preset restrictive condition is a preset length, and the sum of the lengths of all the sentences in the target semantic segment is greater than the preset length; and the step of processing each of the target semantic segments according to the preset restrictive condition to obtain a reference semantic segment corresponding to each of the target semantic segments comprises:

re-segmenting, for any of the target semantic segments, the target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to the target semantic segment if the target semantic segment does not satisfy the preset restrictive condition; and

using the target semantic segment as a reference semantic segment corresponding thereto if the target semantic segment satisfies the preset restrictive condition.

In an optional implementation mode, the step of re-segmenting the target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to the target semantic segment comprises:

using the first sentence in the target semantic segment as a reference sentence;

determining a target sentence sequence, starting from the reference sentence, from the target semantic segment, wherein the sum of the lengths of all the sentences in the target sentence sequence is the maximum and less than the preset length;

affiliating all the sentences in the target sentence sequence to one reference semantic segment; and

if the last sentence in the target sentence sequence is not the last sentence of the target semantic segment, redetermining a reference sentence from the target semantic segment, and returning to the step of determining a target sentence sequence, starting from the reference sentence, from the target semantic segment, until all the reference semantic segments corresponding to the target semantic segment are obtained.

In an optional implementation mode, the step of redetermining a reference sentence from the target semantic segment comprises:

determining a first associated sentence from the target sentence sequence;

determining a second associated sentence from sentences located following the target sentence sequence in the target semantic segment;

determining the reference sentence according to the second associated sentence if the sum of the length of the first associated sentence and the length of the second associated sentence is greater than the preset length, and if not, determining the reference sentence according to the first associated sentence.

In a second aspect, the present disclosure provides a text segmentation apparatus, the apparatus comprising:

an acquisition module configured to acquire a plurality of sentences obtained after sentence-level segmentation of a target text;

a segmentation module configured to slide a preset sliding window multiple times by using the first sentence of the plurality of sentences as a starting point and using the last sentence of the plurality of sentences as an end point, and semantically segment a plurality of target sentences located in the preset sliding window after each sliding operation to obtain at least one semantic segment; and

a clustering module configured to cluster, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained.

In a third aspect, the present disclosure provides an electronic device, comprising a processor and a memory, the memory being configured to store a program, the processor being configured to implement a text segmentation method according to any one of the foregoing implementation modes when executing the program.

In a fourth aspect, the present disclosure provides a computer-readable storage medium, on which a computer program is stored, a text segmentation method according to any one of the foregoing implementation modes being implemented when the computer program is executed by a processor.

In contrast to the prior art, in the text segmentation method, the text segmentation apparatus, the electronic device, and the computer-readable storage medium according to embodiments of the present disclosure, it is disclosed that a preset sliding window is slid for a plurality of sentences, and the target sentences in the preset sliding window are semantically segmented sequentially to ensure the semantic coherence of the segmented target sentences in the preset sliding window, and the semantic segments obtained by the current segmentation and the semantic segments obtained by a previous segmentation are clustered to ensure the semantic coherence between the sentences in the preset sliding window after two successive sliding operations, whereby a reasonable segmentation of the text is finally achieved, and the semantic

coherence of the segmented semantic segments is ensured, thereby improving the accuracy and reliability of a question-answering output to be fed back by RAG.

Brief Description of Drawings

In order to more clearly illustrate the technical solutions of embodiments of the present disclosure, drawings required to be used in the embodiments will be described briefly below. It is understood that the drawings below are merely illustrative of some embodiments of the present disclosure and thus should not be considered as limiting its scope. It will be appreciated by those of ordinary skill in the art that other relevant drawings can be obtained according to these drawings without any inventive effort.

FIG. 1 is a framework diagram of a RAG system according to an embodiment of the present disclosure.

FIG. 2 is an exemplary flowchart of a text segmentation method according to an embodiment of the present disclosure.

FIG. 3 is an exemplary diagram of a process of semantic segmentation using a preset sliding window according to an embodiment of the present disclosure.

FIG. 4 is an exemplary block diagram of a text segmentation apparatus according to an embodiment of the present disclosure.

FIG. 5 is an exemplary block diagram of an electronic device according to an embodiment of the present disclosure.

Reference numerals: 10-electronic device; 11-processor; 12-memory; 13-bus; 100-text segmentation apparatus; 110-acquisition module; 120-segmentation module; 130-clustering module.

Detailed Description of Embodiments

In order to further clarify the objects, technical solutions, and advantages of the embodiments of the present disclosure, the technical solutions of the embodiments of the present disclosure will be described below clearly and completely with reference to the drawings of the embodiments of the present disclosure. It is apparent that the embodiments to be described are some, but not all of the embodiments of the present disclosure. Generally, the components of the embodiments of the present disclosure, as described and illustrated in the figures herein, may be arranged and designed in a wide variety of different configurations.

Thus, the following detailed description of the embodiments of the present disclosure, as represented in the figures, is not intended to limit the scope of the present disclosure as claimed, but is merely representative of selected embodiments of the present disclosure. All the other embodiments obtained by those of ordinary skill in the art in light of the embodiments of the present disclosure without inventive efforts will fall within the scope of the present disclosure as claimed.

It should be noted that similar reference numerals and letters refer to similar items in the following figures, and thus once an item is defined in one figure, it may not be further defined or explained in the following figures.

In the description of the present disclosure, it should be noted that the terms such as “up”, “down”, “inside”, and “outside”, if present, indicate the orientation or positional relationships shown based on the figures, or the orientation or positional relationships in which the disclosed product is conventionally placed in use, and these terms are intended only to facilitate the description of the present disclosure and simplify the description, but not intended to indicate or imply that the referred apparatuses or elements must be in a particular orientation or constructed or operated in the particular orientation, and therefore should not be construed as limiting the present disclosure.

In addition, the terms such as “first” and “second”, if present, are used for distinguishing the description only, and should not be understood as an indication or implication of relative importance.

It should be noted that the features in the embodiments of the present disclosure may be combined with each other in the absence of conflict.

In a prior RAG system, before generating an answer, relevant information is firstly retrieved from an external document, and the retrieved knowledge is inputted into a large model as context, so as to enhance the knowledge coverage range of the large model and improve its accuracy and reliability. Referring to FIG. 1, FIG. 1 is a framework diagram of a RAG system according to an embodiment of the present disclosure. The RAG system comprises the following processing steps:

(1) performing textual segmentation of a document to obtain the segmented text segments;

(2) vectorially storing the segmented text segments to generate a vector database;

(3) when dealing with a specified question, performing vector retrieval in the vector database to retrieve a vector associated with the specified question from the vector database, restoring the retrieved vector to a corresponding text segment (semantic segment in the figure) as context of the specified question, and constructing a prompt for a large model in the RAG system based on the specified question and its context; and

(4) outputting an answer to the question according to the prompt.

In the RAG framework, the segmentation strategy employed in the document segmentation is usually based on a fixed segmentation granularity, and the segmentation granularity is usually associated with an input length limit of a large model. Although such mode with a fixed segmentation granularity can meet the input requirements of the large model, the problem of semantic separation will inevitably occur. Moreover, the input length limit is also different for different large models, and thus the segmentation mode with fixed granularity cannot flexibly adapt to the needs of different scenarios.

In view of this, embodiments of the present disclosure provide a text segmentation method, a text segmentation apparatus, an electronic device, and a computer-readable storage medium, which can reasonably segment a text semantically and ensure the semantic coherence of the segmented semantic segments, which will be described in detail below.

Referring to FIG. 2, FIG. 2 is an exemplary flowchart of a text segmentation method according to an embodiment of the present disclosure. The method comprises the following steps S101 to S103.

In step S101, a plurality of sentences obtained after sentence-level segmentation of a target text is acquired.

In the embodiment of the present disclosure, the target text may be a text extracted from one or more documents which may be Word documents, TXT documents, PDF documents, or the like, or may be a text extracted from a picture after performing character recognition on the picture.

In the embodiment of the present disclosure, the target text may be segmented at the sentence level by using a pretrained model, or the target text may also be segmented at the sentence level according to punctuation marks and line breaks in the target text.

In step S102, a preset sliding window is slid multiple times by using the first sentence of the plurality of sentences as a starting point and using the last sentence of the plurality of sentences as an end point, and a plurality of target sentences located in the preset sliding window are semantically segmented after each sliding operation to obtain at least one semantic segment.

In the embodiment of the present disclosure, the size of the preset sliding window affects the segmentation efficiency and segmentation accuracy. The window size represents the number of sentences that need to be semantically segmented each time. The size of the preset sliding window may be adjusted according to the transaction application scenario. For example, the size of the preset sliding window is 2.

In the embodiment of the present disclosure, the semantic segmentation may be performed using a pretrained language model. The language model includes, but is not limited to, a BERT-NSP model (a deep learning model that processes natural language processing using a Transformer architecture), a BGE model (a text embedding model configured to convert a text into low-dimensional dense vectors so as to perform efficient semantic analysis and retrieval), a SimCSE (Similarity Consensus Semantic Encoding) model (a model configured to generate sentence embeddings), and the like.

In step S103, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments obtained by a previous segmentation are clustered to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained.

In the embodiment of the present disclosure, if the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation are semantically coherent with each other, it is necessary to merge the two semantic segments into one semantic segment, and if not, they are treated as two independent semantic segments, in order to ensure the semantic coherence between the semantic segments obtained by the two segmentations. For example, if the last

semantic segment 1 obtained by the previous segmentation includes sentence a and sentence b, and the first semantic segment 2 obtained by the current segmentation includes sentence b, sentence c, and sentence d, then the semantic segment 1 and the semantic segment 2 are semantically coherent with each other, and thus the semantic segment 1 and the semantic segment 2 may be merged into one semantic segment.

In the embodiment of the present disclosure, as an optional implementation, one clustering operation may be performed following one sliding operation, or semantic segments obtained by semantic segmentation after each sliding operation may be clustered successively after multiple sliding operations are performed. Both eventually enable obtainment of a semantic segment to which each sentence of a plurality of sentences belongs.

In the above method according to the embodiment of the present disclosure, the target sentences in the preset sliding window are semantically segmented sequentially to ensure the semantic coherence of the segmented target sentences in the preset sliding window, and the semantic segments obtained by the current segmentation and the semantic segments obtained by the previous segmentation are clustered to ensure the semantic coherence between the sentences in the preset sliding window after two successive sliding operations, thereby achieving a reasonable segmentation of the text.

In an optional implementation, in order to semantically segment a plurality of target sentences, they may be segmented by determining whether every two adjacent target sentences are semantically continuous with each other. An optional implementation is performed as follows.

Firstly, the plurality of target sentences are inputted into a pretrained text segmentation model to obtain the semantic continuity of sentence pairs composed of every two adjacent target sentences.

In the embodiment of the present disclosure, the text segmentation model may be a model based on BERT-NSP, BGE, or SimCSE (Similarity Consensus Semantic Encoding). Taking the text segmentation model based on BERT-NSP as an example, phrases or sentences may be labelled as being misplaced according to the content of a document during construction of a training dataset. Namely, when processing the document, those sentences inconsistent with the overall content of the document or having errors are identified and labelled, so as to automatically correct the errors in the document based on the labelling of misplacements, improve the accuracy of extraction of information from the document, and accurately evaluate the overall quality and consistency of the document.

In addition, in the case where there are relatively scarce samples between paragraphs in the dataset, an oversampling strategy may also be employed. Namely, the number of samples of each class are balanced by increasing the number of samples of a minority class. The common oversampling strategy includes, but is not limited to, random oversampling and interpolation between the samples of the minority class to generate a new combined sample so as to ensure the balance of samples of training data, thereby improving the generalization capability of the model.

Secondly, for any target sentence pair, if the semantic continuity of the target sentence pair is greater than a preset value, the two target sentences in the target sentence pair are affiliated to the same semantic segment, and if not, the two target sentences in the target sentence pair are affiliated to different semantic segments.

In the embodiment of the present disclosure, as an optional implementation, the semantic continuity obtained by the text segmentation model may be a probability value of the semantic continuity between two target sentences, the two target sentences are considered to belong to the same semantic segment if it is greater than a preset value, and if not, the two target sentences are considered to belong to different semantic segments. As another optional implementation, the semantic continuity obtained by the text segmentation model may be a character string composed of 0 and/or 1, where 0 denotes that corresponding two target sentences are semantically discontinuous with each other, and 1 denotes that corresponding two target sentences are semantically continuous with each other. For example, the plurality of target sentences are sentence a, sentence b, sentence c, and sentence d, sentence a and sentence b form sentence pair 1, sentence b and sentence c form sentence pair 2, and sentence c and sentence d form sentence pair 3. If the character string outputted by the text segmentation model is 011, then "0" denotes the semantic continuity of sentence pair 1, representing that sentence a and sentence b are semantically discontinuous with each other, the first "1" denotes the semantic continuity of sentence pair 2, representing that sentence b and sentence c are semantically continuous with each other, and the second "1" denotes the semantic continuity of sentence pair 3, namely, representing that sentence c and sentence d are semantically continuous with each other.

Finally, each sentence pair is sequentially treated as the target sentence pair to obtain at least one semantic segment after semantic segmentation of the plurality of target sentences.

In the embodiment of the present disclosure, semantically continuous sentences in the plurality of target sentences are classified into the same semantic segment, and semantically discontinuous sentences are classified into different semantic segments, so as to obtain at least one semantic segment after semantic segmentation of the plurality of target sentences.

In an optional implementation mode, in order to ensure the semantic coherence between the semantic segments obtained by successive segmentations, the last target sentence in the preset sliding window in the previous sliding operation is used as the first target sentence in the preset sliding window after the current sliding operation. Namely, the first target sentence in the preset sliding window after the current sliding operation is the same as the last target sentence in the preset sliding window in the previous sliding operation. On this basis, an implementation of the clustering of the semantic segments obtained by the current segmentation and the semantic segments obtained by the previous segmentation is performed as follows.

Firstly, the semantic continuity of the first sentence pair in the preset sliding window after the current sliding operation is acquired.

In the embodiment of the present disclosure, the semantic continuity of the first sentence pair may be obtained by inputting two sentences in the first sentence pair into the text segmentation model described above.

Secondly, the semantic similarity between the two target sentences in the first sentence pair is calculated.

In the embodiment of the present disclosure, as an optional implementation, the semantic similarity may be obtained by inputting the two target sentences into a BGE, or M3E (Multi-Modal Multi-Task Embedding) model or the like. As another implementation, feature vectors of the two target sentences are extracted respectively, then cosine similarity between their feature vectors are calculated, and the calculated cosine similarity is used as the semantic similarity.

Thirdly, a semantic association score of the first sentence pair is calculated according to the semantic continuity and the semantic similarity between the two target sentences in the first sentence pair.

In the embodiment of the present disclosure, a continuity weight and a similarity weight may be set for the semantic continuity and the semantic similarity, respectively, where the semantic association score = semantic continuity × continuity weight + semantic similarity × similarity weight. The semantic continuity weight and the semantic similarity weight may be set according to the actual application scenario and the reliability of its corresponding model, or its matching degree with the actual application scenario. For example, the continuity weight is set to 0.7, and the similarity weight is set to 0.3.

Finally, if the semantic association score is greater than a preset score, the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation are merged into one semantic segment, and if not, the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation are treated as different semantic segments.

In the embodiment of the present disclosure, the preset score may be set according to actual needs. For example, the preset score is 0.5.

It should be noted that, when allowed by the semantic segmentation efficiency, the plurality of target sentences located in the preset sliding window may be semantically segmented using the semantic association score after each sliding operation to obtain at least one semantic segment. In other words, another implementation for semantically segmenting the plurality of target sentences located in the preset sliding window to obtain at least one semantic segment may be performed by:

acquiring semantic continuity of a sentence pair composed of every two adjacent target sentences;

calculating semantic similarity of the sentence pair composed of every two adjacent target sentences;

calculating a semantic association score of the sentence pair composed of every two adjacent target sentences according to the semantic continuity and the semantic similarity between the two target sentences in the sentence pair composed of every two adjacent target sentences; and

if the semantic association score is greater than a preset score, affiliating the target sentences in the corresponding sentence pair to the same semantic segment, and if not, affiliating them to different semantic segments.

Further, in order to improve the efficiency of semantic segmentation, the semantic similarity of a semantic pair(s) to be confirmed may be calculated only for a semantic pair(s) to be confirmed, in the plurality of target sentences, that has a semantic continuity less than or equal to the preset value, and a semantic association score of the semantic pair to be confirmed is calculated according to the semantic continuity and semantic similarity of the semantic pair to be confirmed. If the semantic association score is greater than the preset value, the target sentences in the semantic pair to be confirmed are affiliated to the same semantic segment, and if not, they are affiliated to different semantic segments. Since the semantic association score is calculated, only for the semantic pair(s) to be confirmed that has a semantic continuity less than or equal to the preset value, to confirm whether the target sentences included therein are affiliated to the same semantic segment, the amount of computation is reduced to a certain extent, and the efficiency of semantic segmentation is improved to a certain extent as compared with the mode in which the semantic association score of each sentence pair in the target sentences is calculated.

In order to more intuitively show the process of semantic segmentation of a plurality of sentences during sliding of the preset sliding window, reference can be made to FIG. 3. FIG. 3 is an exemplary diagram of a process of semantic segmentation using a preset sliding window according to an embodiment of the present disclosure. Semantic segmentation is performed for sentence 1, sentence 2, and sentence 3 in the preset sliding window to obtain: a semantic segment [sentence 1, sentence 2] and a semantic segment [sentence 3], the preset sliding window is slid and semantic segmentation is performed for sentence 3, sentence 4, and sentence 5 in the preset sliding window to obtain: a semantic segment [sentence 3, sentence 4, sentence 5], and the semantic segment [sentence 3, sentence 4, sentence 5] obtained currently and the semantic segment [sentence 1, sentence 2] and semantic segment [sentence 3] obtained previously are clustered to obtain a semantic segment [sentence 1, sentence 2] and a semantic segment [sentence 3, sentence 4, sentence 5].

In order to show the effect of the semantic segmentation mode according to the embodiment of the present disclosure, in the embodiment of the present disclosure, sentences consisting of manually labeled 19 segments are segmented in three segmentation modes, and the segmentation results of the three segmentation modes are evaluated and compared. Mode 1: the segmentation is performed according to a fixed number of characters or words. Mode 2: in the embodiment of the present disclosure, the segmentation is performed using a preset sliding window and a BERT-NSP model. Mode 3: in the embodiment of the present disclosure, the segmentation is performed using a preset sliding window and a BERT-NSP+M3E model.

Segmentation result of mode 1: 18 segments, including two correctly segmented segments; P=11.11%, R=94.73%, F1=19.87%;

Segmentation result of mode 2: 16 segments, including 13 correctly segmented segments, P=81.25%, R=84.21%, F1=82.70%;

Segmentation result of mode 3: 18 segments, including 16 correctly segmented segments, P=88.88%, R=94.73%, F1=91.71%;

where P denotes an accuracy rate, which is calculated by the number of correctly segmented segments / the total number of segmented segments; R denotes a recall rate, which is calculated by the number of correctly segmented segments / the total number of segments that should be segmented; and F1 denotes a score of the segmentation result, which is calculated by 2 × accuracy rate × recall rate / (accuracy rate + recall rate).

In the embodiment of the present disclosure, after a target semantic segment to which each target sentence in the preset sliding window after the previous sliding operation and after the current sliding operation belongs is obtained, it is further necessary to process the target semantic segment in accordance with the input requirements of a required large model, in order to enable the segmented semantic segments to flexibly adapt to the input requirements of different large models. A processing mode is provided in the embodiment of the present disclosure.

Firstly, a preset restrictive condition is acquired, where the preset restrictive condition is configured to restrict semantic segments to be inputted into a large model in a retrieval-augmented generation (RAG) system.

In the embodiment of the present disclosure, the preset restrictive condition may be the maximum length of a single semantic segment to be inputted to a large model. For example, the maximum length of each semantic segment does not exceed 500 characters or words. Alternatively, the preset restrictive condition may be the degree of overlap between two adjacent semantic segments to be inputted to the large model. For example, the first semantic segment is from the first sentence to the 10th sentence, and the second semantic segment is from the 6th sentence to the 15th sentence.

It should be noted that the preset restrictive condition may also be a combination of the maximum length and the degree of overlap, namely, it restricts both the length of a single semantic segment to be inputted to the large model, and the degree of overlap between two adjacent semantic segments.

Secondly, each target semantic segment is processed according to the preset restrictive condition to obtain a reference semantic segment corresponding to each target semantic segment, and each reference semantic segment is stored into a vector database in a vectorized form, so that when the large model receives input data, a semantic segment vector related to the input data is retrieved from the vector database, and feedback is performed based on the input data and the retrieved semantic segment vector.

In the embodiment of the present disclosure, the reference semantic segment is a semantic segment that satisfies the preset restrictive condition. One target semantic segment may correspond to one reference semantic segment, or one target semantic segment may correspond to a plurality of reference semantic segments. When one target semantic segment corresponds to a plurality of reference semantic segments, the plurality of reference semantic segments may partially overlap with each other, in order to ensure the semantic coherence between the plurality of reference semantic segments.

In an optional implementation mode, for a large model with limited input length, in order to meet the length requirement of the large model while maintaining the semantic coherence between sentences as much as possible, an implementation for processing

the target semantic segment to obtain a reference semantic segment that satisfies the input length requirement of the large model is provided in the embodiment of the present disclosure.

Firstly, for any target semantic segment, if the target semantic segment does not satisfy the preset restrictive condition, the target semantic segment is re-segmented according to the preset restrictive condition to obtain a reference semantic segment corresponding to the target semantic segment.

Secondly, if the target semantic segment satisfies the preset restrictive condition, the target semantic segment is used as a reference semantic segment corresponding thereto.

In the embodiment of the present disclosure, the preset restrictive condition is a preset length. If the target semantic segment satisfies the preset restrictive condition, meaning that the sum of the lengths of all the sentences in the target semantic segment is less than or equal to the preset length, then the target semantic segment may not be processed, and the target semantic segment is a reference semantic segment. If not, it is meant that the sum of the lengths of all the sentences in the target semantic segment is greater than the preset length, then it is necessary to re-segment the target semantic segment, so that the segmented semantic segments satisfy the preset restrictive condition. An implementation of the re-segmentation is performed by:

(1) using the first sentence in the target semantic segment as a reference sentence;

(2) determining a target sentence sequence, starting from the reference sentence, from the target semantic segment, wherein the sum of the lengths of all the sentences in the target sentence sequence is the maximum and less than the preset length;

(3) affiliating all the sentences in the target sentence sequence to one reference semantic segment; and

(4) if the last sentence in the target sentence sequence is not the last sentence of the target semantic segment, redetermining a reference sentence from the target semantic segment, and returning to the step (2) until all the reference semantic segments corresponding to the target semantic segment are obtained.

In an optional implementation mode, in order to maximally ensure the semantic coherence between the finally obtained reference semantic segments during re-segmentation, an implementation for determining the reference sentence is provided in the embodiment of the present disclosure.

Firstly, a first associated sentence is determined from the target sentence sequence.

Secondly, a second associated sentence is determined from sentences located following the target sentence sequence in the target semantic segment.

Finally, if the sum of the length of the first associated sentence and the length of the second associated sentence is greater than the preset length, the reference sentence is determined according to the second associated sentence, and if not, the reference sentence is determined according to the first associated sentence.

In the embodiment of the present disclosure, the determination of the reference sentence according to the second associated sentence may be performed by using the first sentence of the second associated sentence as the reference sentence, or using the first sentence of a number of subsequent sentences included in the first associated sentence as the reference sentence, as long as the sum of the length of a number of subsequent sentences of the first associated sentence and the length of the second associated sentence is ensured to be less than or equal to the preset length.

In the embodiment of the present disclosure, in the embodiment of the present disclosure, the number of each of the first associated sentences and the second associated sentences may be one or more, depending on the maximum number of overlaps allowed during segmentation. The first associated sentence may be a preset number of subsequent sentences located in the target sentence sequence. As an implementation, the preset number may be fixed depending on the maximum number of overlaps allowed during segmentation. For example, the preset number is set to 2. As another implementation, the preset number may also be variable, as long as it can be satisfied that the sum of the length of the first associated sentence(s) and the length of the second associated sentence(s) is less than or equal to the preset length.

The embodiment of the present disclosure is described exemplarily by taking two first associated sentences and one second associated sentence as an example. For example, a target semantic segment includes the following sentences:

[sentence 1, sentence 2, sentence 3, sentence 4, sentence 5, sentence 6, sentence 7]

If sentence 1 is used as the reference sentence, and the sum of the lengths of sentence 1, sentence 2, and sentence 3 is the maximum and less than the preset length, then the target sentence sequence is (sentence 1, sentence 2, sentence 3), and sentence 1, sentence 2, and sentence 3 belong to reference semantic segment 1.

Sentence 2 and sentence 3 are used as the first associated sentences, and sentence 4 is used as the second associated sentence. If the sum of the lengths of sentence 2, sentence 3, and sentence 4 is less than the preset length, sentence 2 is used as the reference sentence.

If the sum of the lengths of sentence 2, sentence 3, sentence 4, and sentence 5 is the maximum and less than the preset length, the target sentence sequence is (sentence 2, sentence 3, sentence 4, sentence 5), and sentence 2, sentence 3, sentence 4, and sentence 5 belong to reference semantic segment 2.

Sentence 4 and sentence 5 are used as the first associated sentences, and sentence 6 is used as the second associated sentence. If the sum of the lengths of sentence 4, sentence 5, and sentence 6 is greater than the preset length, sentence 6 may be used as the reference sentence. If the sum of the lengths of sentence 5 and sentence 6 is less than the preset length, sentence 5 may also be used as the reference sentence. In this example, sentence 5 is used as the reference sentence.

If the sum of the lengths of sentence 5, sentence 6, and sentence 7 is less than the preset length, the target sentence sequence is (sentence 5, sentence 6, sentence 7),

and sentence 5, sentence 6, and sentence 7 belong to reference semantic segment 3.

Thus, the target semantic segment is re-segmented into three reference semantic segments: reference semantic segment 1 to reference semantic segment 3.

In the embodiment of the present disclosure, the preset restrictive condition may be the number of overlaps, in addition to the preset length. If the number of overlaps is 2, a target semantic segment [sentence 1, sentence 2, sentence 3, sentence 4, sentence 5] may be re-segmented into two reference semantic segments (sentence 1, sentence 2, sentence 3, sentence 4) and (sentence 3, sentence 4, sentence 5) which jointly contain two overlapping sentences: sentence 3 and sentence 4, or may be re-segmented into (sentence 1, sentence 2, sentence 3), (sentence 2, sentence 3, sentence 4), and (sentence 3, sentence 4, sentence 5), depending on actual needs.

In order to execute the corresponding steps in the above embodiment and in each possible implementation mode, an implementation of a text segmentation apparatus 100 is given below, respectively. Referring to FIG. 4, FIG. 4 is a schematic block diagram of a text segmentation apparatus according to an embodiment of the present disclosure. It should be noted that the text segmentation apparatus 100 according to the present disclosure has the same basic principle and achieves the same technical effects as the above corresponding embodiment, and part of the embodiment of the present disclosure is not mentioned or stated for the sake of a brief description.

The text segmentation apparatus 100 comprises an acquisition module 110, a segmentation module 120, and a clustering module 130.

The acquisition module 110 is configured to acquire a plurality of sentences obtained after sentence-level segmentation of a target text.

The segmentation module 120 is configured to slide a preset sliding window multiple times by using the first sentence of the plurality of sentences as a starting point and using the last sentence of the plurality of sentences as an end point, and semantically segment a plurality of target sentences located in the preset sliding window after each sliding operation to obtain at least one semantic segment.

The clustering module 130 is configured to cluster, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained.

In an optional implementation mode, the segmentation module 120 is configured to:

input the plurality of target sentences into a pretrained text segmentation model to obtain the semantic continuity of sentence pairs composed of every two adjacent target sentences;

affiliate, for any target sentence pair, the two target sentences in the target sentence pair to the same semantic segment if the semantic continuity of the target sentence

pair is greater than a preset value, and if not, affiliate the two target sentences in the target sentence pair to different semantic segments; and

sequentially treat each sentence pair as the target sentence pair to obtain at least one semantic segment after semantic segmentation of the plurality of target sentences.

In an optional implementation mode, the first target sentence in the preset sliding window after the current sliding operation is the same as the last target sentence in the preset sliding window in the previous sliding operation, and every two adjacent target sentences in the preset sliding window form a sentence pair.

For the semantic segments obtained by each segmentation, the clustering module 130 is configured to:

acquire the semantic continuity of the first sentence pair in the preset sliding window after the current sliding operation;

calculate the semantic similarity between the two target sentences in the first sentence pair;

calculate a semantic association score of the first sentence pair according to the semantic continuity and the semantic similarity between the two target sentences in the first sentence pair;

merge the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation are merged into one semantic segment if the semantic association score is greater than a preset score, and if not, treat the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation as different semantic segments.

In an optional implementation mode, the clustering module 130 is further configured to:

acquire a preset restrictive condition, where the preset restrictive condition is configured to restrict semantic segments to be inputted into a large model in a retrieval-augmented generation (RAG) system;

process each target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to each target semantic segment, and store each reference semantic segment into a vector database in a vectorized form, so that when the large model receives input data, a semantic segment vector related to the input data is retrieved from the vector database, and feedback is performed based on the input data and the retrieved semantic segment vector.

In an optional implementation mode, the preset restrictive condition is a preset length, and the sum of the lengths of all the sentences in the target semantic segment is greater than the preset length. When configured to process each target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to each target semantic segment, the clustering module 130 is configured to:

re-segment, for any target semantic segment, the target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to the target semantic segment if the target semantic segment does not satisfy the preset restrictive condition; and

use the target semantic segment as a reference semantic segment corresponding thereto if the target semantic segment satisfies the preset restrictive condition.

In an optional implementation mode, when configured to re-segment the target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to the target semantic segment, the clustering module 130 is configured to:

use the first sentence in the target semantic segment as a reference sentence;

determine a target sentence sequence, starting from the reference sentence, from the target semantic segment, wherein the sum of the lengths of all the sentences in the target sentence sequence is the maximum and less than the preset length;

affiliate all the sentences in the target sentence sequence to one reference semantic segment; and

if the last sentence in the target sentence sequence is not the last sentence of the target semantic segment, redetermine a reference sentence from the target semantic segment, and return to the step of determining a target sentence sequence, starting from the reference sentence, from the target semantic segment, until all the reference semantic segments corresponding to the target semantic segment are obtained.

In an optional implementation mode, when configured to redetermine a reference sentence from the target semantic segment, the clustering module 130 is configured to:

determine a first associated sentence from the target sentence sequence;

determine a second associated sentence from sentences located following the target sentence sequence in the target semantic segment;

determine the reference sentence according to the second associated sentence if the sum of the length of the first associated sentence and the length of the second associated sentence is greater than the preset length, and if not, determine the reference sentence according to the first associated sentence.

An embodiment of the present disclosure further provides a schematic block diagram of an electronic device 10. The electronic device 10 implements the text segmentation method of the foregoing embodiment. Referring to FIG. 5, FIG. 5 is a schematic block diagram of an electronic device 10 according to an embodiment of the present disclosure. The electronic device 10 comprises a processor 11, a memory 12, and a bus 13, and the processor 11 and the memory 12 are connected by the bus 13.

The processor 11 may be an integrated circuit chip having the capability of processing signals. During the implementation, each step of the text segmentation method of the above embodiment may be completed by an integrated logic circuit of

hardware in the processor 11 or by an instruction in the form of software. The above-mentioned processor 11 may be a general-purpose processor, including a CPU (central processing unit), an NP (network processor), and the like, or may be a DSP (digital signal processor), an ASIC (application specific integrated circuit), an FPGA (field programmable logic gate array) or other programmable logic device, a discrete gate or transistor logic device, or a discrete hardware component.

The memory 12 is configured to store a program for implementing the text segmentation method. The program may be a software functional module stored in the memory 12 in the form of software or firmware or solidified in an operating system (OS) of the electronic device 10.

After receiving an execution instruction, the processor 11 executes the program to implement the text segmentation method of the foregoing embodiment.

An embodiment of the present disclosure provides a computer-readable storage medium on which a computer program is stored, and a text segmentation method as described above is implemented when the computer program is executed by a processor.

In summary, the embodiments of the present disclosure provide a text segmentation method, a text segmentation apparatus, an electronic device, and a computer-readable storage medium. The method comprises: acquiring a plurality of sentences obtained after sentence-level segmentation of a target text; sliding a preset sliding window multiple times by using the first sentence of the plurality of sentences as a starting point and using the last sentence of the plurality of sentences as an end point, and semantically segmenting a plurality of target sentences located in the preset sliding window after each sliding operation to obtain at least one semantic segment; and clustering, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained. Compared with the prior art, the embodiments of the present disclosure have at least the following advantages. (1) A preset sliding window is slid for a plurality of sentences, and the target sentences in the preset sliding window are semantically segmented sequentially to ensure the semantic coherence of the segmented target sentences in the preset sliding window, and the semantic segments obtained by the current segmentation and the semantic segments obtained by a previous segmentation are clustered to ensure the semantic coherence between the sentences in the preset sliding window after two successive sliding operations, whereby a reasonable segmentation of the text is finally achieved, and the semantic coherence of the segmented semantic segments is ensured, thereby improving the accuracy and reliability of a question-answering output to be fed back by RAG. (2) The semantic continuity and semantic similarity between the sentences in the preset sliding window after two successive sliding operations are used to accurately judge the semantic coherence therebetween so as to improve the reasonability of the semantic segmentation. (3) The present disclosure is more versatile and economical and achieves the purpose of reducing costs and increasing efficiency, because it ensures the semantic coherence between sentences in the semantically segmented semantic segments, and reduces the subsequent hardware

resource requirements for a large model of RAG. (4) A preset restrictive condition configured to restrict semantic segments to be inputted into a large model in a RAG system can be taken into consideration simultaneously during segmentation, and the semantic segments obtained by segmentation are processed according to the preset restrictive condition, so that the final segmentation result can meet the needs of the large model in the RAG system while maximally ensuring the semantic coherence, thereby achieving the effect of being flexibly adaptable to the requirements of different large models.

The above description is only illustrative of various implementation modes of the present disclosure, but the scope of protection of the present disclosure is not limited thereto. Any modifications or substitutions that can be readily envisaged by those skilled in the art within the scope of the technologies disclosed in the present disclosure shall be encompassed by the scope of protection of the present disclosure. Therefore, the scope of protection of the present disclosure shall be defined by the scope of protection of the claims.

Industrial Applicability

The embodiments of the present disclosure provide a text segmentation method, a text segmentation apparatus, an electronic device, and a computer-readable storage medium. The method comprises: acquiring a plurality of sentences obtained after sentence-level segmentation of a target text; sliding a preset sliding window multiple times by using the first sentence of the plurality of sentences as a starting point and using the last sentence of the plurality of sentences as an end point, and semantically segmenting a plurality of target sentences located in the preset sliding window after each sliding operation to obtain at least one semantic segment; and clustering, for semantic segments obtained by each segmentation, the semantic segments obtained by the current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained. The embodiments of the present disclosure have at least the following advantages. (1) A preset sliding window is slid for a plurality of sentences, and the target sentences in the preset sliding window are semantically segmented sequentially to ensure the semantic coherence of the segmented target sentences in the preset sliding window, and the semantic segments obtained by the current segmentation and the semantic segments obtained by a previous segmentation are clustered to ensure the semantic coherence between the sentences in the preset sliding window after two successive sliding operations, whereby a reasonable segmentation of the text is finally achieved, and the semantic coherence of the segmented semantic segments is ensured, thereby improving the accuracy and reliability of a question-answering output to be fed back by RAG. (2) The semantic continuity and semantic similarity between the sentences in the preset sliding window after two successive sliding operations are used to accurately judge the semantic coherence therebetween so as to improve the reasonability of the semantic segmentation. (3) The present disclosure is more versatile and economical and achieves the purpose of reducing costs and increasing efficiency, because it ensures the semantic coherence between sentences in the semantically segmented semantic segments, and reduces the subsequent hardware resource requirements for a large model of RAG. (4) A preset restrictive condition configured to restrict semantic

segments to be inputted into a large model in a RAG system can be taken into consideration simultaneously during segmentation, and the semantic segments obtained by segmentation are processed according to the preset restrictive condition, so that the final segmentation result can meet the needs of the large model in the RAG system while maximally ensuring the semantic coherence, thereby achieving the effect of being flexibly adaptable to the requirements of different large models.

CLAIMS

Claim 1: A text segmentation method, the method comprising:

acquiring a plurality of sentences obtained after sentence-level segmentation of a target text;

sliding a preset sliding window multiple times by using a first sentence of the plurality of sentences as a starting point and using a last sentence of the plurality of sentences as an end point, and semantically segmenting a plurality of target sentences located in the preset sliding window after each sliding operation to obtain at least one semantic segment; and

clustering, for semantic segments obtained by each segmentation, semantic segments obtained by a current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained.

Claim 2: The text segmentation method according to claim 1, wherein the step of semantically segmenting a plurality of target sentences located in the preset sliding window to obtain at least one semantic segment comprises:

inputting the plurality of target sentences into a pretrained text segmentation model to obtain semantic continuity of sentence pairs composed of every two adjacent target sentences;

affiliating, for any target sentence pair, two target sentences in the target sentence pair to the same semantic segment if a semantic continuity of the target sentence pair is greater than a preset value, and if not, affiliating the two target sentences in the target sentence pair to different semantic segments; and

sequentially treating each sentence pair as the target sentence pair to obtain at least one semantic segment after semantic segmentation of the plurality of target sentences.

Claim 3: The text segmentation method according to claim 1, wherein the first target sentence in the preset sliding window after the current sliding operation is the same as the last target sentence in the preset sliding window in the previous sliding operation, and every two adjacent target sentences in the preset sliding window form a sentence pair;

the step of clustering, for semantic segments obtained by each segmentation, semantic segments obtained by a current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs comprises:

acquiring semantic continuity of a first sentence pair in the preset sliding window after the current sliding operation;

calculating semantic similarity between two target sentences in the first sentence pair;

calculating a semantic association score of the first sentence pair according to the semantic continuity and the semantic similarity between the two target sentences in the first sentence pair; and

merging the last semantic segment obtained by a previous segmentation and the first semantic segment obtained by a current segmentation into one semantic segment if the semantic association score is greater than a preset score, and if not, treating the last semantic segment obtained by the previous segmentation and the first semantic segment obtained by the current segmentation as different semantic segments.

Claim 4: The text segmentation method according to claim 3, wherein the step of calculating semantic similarity between two target sentences in the first sentence pair comprises:

extracting feature vectors of the two target sentences, respectively;

calculating cosine similarity between the feature vectors of the two target sentences, and using the calculated cosine similarity as the semantic similarity.

Claim 5: The text segmentation method according to claim 3, wherein the step of calculating semantic similarity between two target sentences in the first sentence pair comprises:

inputting the two target sentences into a multimodal multi-task embedding model to obtain the semantic similarity between the two target sentences.

Claim 6: The text segmentation method according to claim 3, wherein the semantic association score of the first sentence pair is calculated by a following formula:

semantic association score = semantic continuity × continuity weight + semantic similarity × similarity weight.

Claim 7: The text segmentation method according to claim 1, wherein the step of semantically segmenting a plurality of target sentences located in the preset sliding window to obtain at least one semantic segment comprises:

acquiring semantic continuity of a sentence pair composed of every two adjacent target sentences;

calculating semantic similarity of the sentence pair composed of every two adjacent target sentences;

calculating a semantic association score of the sentence pair composed of every two adjacent target sentences according to the semantic continuity and the semantic similarity between the two target sentences in the sentence pair composed of every two adjacent target sentences; and

if the semantic association score is greater than a preset score, affiliating target sentences in a corresponding sentence pair to the same semantic segment, and if not, affiliating them to different semantic segments.

Claim 8: The text segmentation method according to any one of claims 1 to 7, wherein after the step of clustering, for semantic segments obtained by each segmentation, semantic segments obtained by a current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, the method further comprises:

acquiring a preset restrictive condition, where the preset restrictive condition is configured to restrict semantic segments to be inputted into a large model in a retrieval-augmented generation (RAG) system; and

processing each of the target semantic segments according to the preset restrictive condition to obtain a reference semantic segment corresponding to each of the target semantic segments, and storing each of the reference semantic segments into a vector database in a vectorized form, so that when the large model receives input data, a

semantic segment vector related to the input data is retrieved from the vector database, and feedback is performed based on the input data and the retrieved semantic segment vector.

Claim 9: The text segmentation method according to claim 8, wherein the preset restrictive condition comprises a maximum length of a single semantic segment to be inputted into a large model and/or a degree of overlap between two adjacent semantic segments to be inputted into the large model.

Claim 10: The text segmentation method according to claim 8, wherein the preset restrictive condition is a preset length, and a sum of lengths of all sentences in the target semantic segment is greater than the preset length; and the step of processing each of the target semantic segments according to the preset restrictive condition to obtain a reference semantic segment corresponding to each of the target semantic segments comprises:

re-segmenting, for any of the target semantic segments, the target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to the target semantic segment if the target semantic segment does not satisfy the preset restrictive condition; and

using the target semantic segment as a reference semantic segment corresponding thereto if the target semantic segment satisfies the preset restrictive condition.

Claim 11: The text segmentation method according to claim 10, wherein the step of re-segmenting the target semantic segment according to the preset restrictive condition to obtain a reference semantic segment corresponding to the target semantic segment comprises:

using a first sentence in the target semantic segment as a reference sentence;

determining a target sentence sequence, starting from the reference sentence, from the target semantic segment, wherein a sum of lengths of all sentences in the target sentence sequence is the maximum and less than the preset length;

affiliating all the sentences in the target sentence sequence to one reference semantic segment; and

if a last sentence in the target sentence sequence is not a last sentence of the target semantic segment, redetermining a reference sentence from the target semantic segment, and returning to the step of

determining a target sentence sequence, starting from the reference sentence, from the target semantic segment, until all the reference semantic segments corresponding to the target semantic segment are obtained.

Claim 12: The text segmentation method according to claim 11, wherein the step of redetermining a reference sentence from the target semantic segment comprises:

determining a first associated sentence from the target sentence sequence;

determining a second associated sentence from sentences located following the target sentence sequence in the target semantic segment; and

determining the reference sentence according to the second associated sentence if a sum of a length of the first associated sentence and a length of the second associated sentence is greater than the preset length, and if not, determining the reference sentence according to the first associated sentence.

Claim 13: A text segmentation apparatus, the apparatus comprising:

an acquisition module configured to acquire a plurality of sentences obtained after sentence-level segmentation of a target text;

a segmentation module configured to slide a preset sliding window multiple times by using a first sentence of the plurality of sentences as a starting point and using a last sentence of the plurality of sentences as an end point, and semantically segment a plurality of target sentences located in the preset sliding window after each sliding operation to obtain at least one semantic segment; and

a clustering module configured to cluster, for semantic segments obtained by each segmentation, semantic segments obtained by a current segmentation and semantic segments obtained by a previous segmentation to obtain a target semantic segment to which each of the target sentences in the preset sliding window after the previous sliding operation and after the current sliding operation belongs, until a semantic segment to which each sentence of the plurality of sentences belongs is obtained.

Claim 14: An electronic device, comprising a processor and a memory, the memory being configured to store a program, the processor being

configured to implement the text segmentation method according to any one of claims 1 to 12 when executing the program.

Claim 15: A computer-readable storage medium, on which a computer program is stored, the text segmentation method according to any one of claims 1 to 12 being implemented when the computer program is executed by a processor.