

CLAIMS

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

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

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

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

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

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

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

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

5

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

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

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

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

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

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

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

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

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

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