

CLAIMS

1. A machine learning-based language correction system comprising:
a correction model learning unit that performs machine learning on a plurality of data sets including ungrammatical sentence data and error-free grammatical sentence data respectively corresponding to the ungrammatical sentence data to generate a correction mode for detecting grammatical sentence data corresponding to ungrammatical sentence data to be corrected; and
a language correction unit that generates, for a sentence to be corrected, a corresponding corrected sentence by using the correction model generated by the correction model learning unit, and indicates and outputs the corrected parts together with the generated corrected sentence,

wherein the correction model learning unit includes:

a pre-processing part that performs a filtering task into a monolingual sentence and a data purification and normalization task by performing language detection on the ungrammatical sentence data;

a learning processing part that performs a supervised learning data labeling task, a machine learning data expansion task, and a parallel data construction task for machine learning, with respect to a plurality of data sets filtered by the pre-processing part;

a correction learning part that generates the corresponding correction model by performing supervised learning-based machine learning on the data sets processed by the learning processing part; and
a post-processing part that outputs errors and error category information through tag additional information added during the supervised learning data labeling task in the learning processing part and then removes the corresponding tag additional information, and
wherein the parallel data construction task for machine learning in the learning processing part includes a task of constructing parallel data using a parallel corpus formed by pairing ungrammatical sentences unnecessary for correction and corresponding grammatical sentences.

~~2. The language correction system of claim 1, wherein the correction model learning unit includes:~~
~~a pre-processing part that performs a filtering task into a monolingual sentence and a data purification and normalization task by performing language detection on the ungrammatical sentence data;~~
~~a learning processing part that performs a supervised learning data labeling task, a machine learning data expansion task, and a parallel data construction task for machine learning, with respect to a plurality of data sets filtered by the pre-processing part ;~~

~~a correction learning part that generates the corresponding correction model by performing supervised learning-based machine learning on the data sets processed by the learning processing part; and~~
~~a first post-processing part that outputs errors and error category information through tag additional information added during the supervised learning data labeling task in the learning processing part and then removes the corresponding tag additional information.~~

32. The machine learning-based language correction system of claim 21, wherein the machine learning data expansion task in the learning processing part includes a data expansion task using letters formed from surrounding typing error characters around an in-position of a keyboard for typing letters included in the ungrammatical sentence data.

~~4. The machine learning-based language correction system of claim 2, wherein the parallel data construction task for machine learning in the learning processing part includes a task of constructing parallel data using a parallel corpus formed by paring ungrammatical sentences unnecessary for correction and corresponding grammatical sentences.~~

53. The machine learning-based language correction system of claim 21, wherein the correction learning part provides an error occurrence probability value, for a learning result in the supervised learning-based

machine learning, as attention weight information between the ungrammatical sentence data and the grammatical sentence data.

64.The machine learning-based language correction system of claim 21, further comprising:

a translation engine for translating input sentences into a preset language, wherein

the pre-processing part marks words, which are unregistered in a dictionary used by the translation engine, by using a preset marker while performing a translation on a large amount of ungrammatical sentence data in the data sets through the translation engine, completes the translation on the large amount of ungrammatical sentence data, and then performs preliminary correction of extracting the words marked by the preset marker to collectively correct the words into error-free words.

75.The machine learning-based language correction system of claim 64, wherein the pre-processing part checks a frequency while extracting the words marked by the preset marker, and aligns the words marked by the preset marker based on the checked frequency to collectively correct the words into error-free words.

86.The machine learning-based language correction system of claim 1, wherein the language correction unit includes:

a pre-processing part that performs pre-process of performing a sentence segmentation, for sentences to be corrected, in a unit of sentence, and tokenizing the segmented sentences;

an error sentence detection part for classifying an error sentence and a non-error sentence by using a binary classifier on the sentence to be corrected that has been pre-processed by the pre-processing part;

a spelling correction part for correcting a spelling error on the sentence to be corrected when the sentence to be corrected is classified as the error sentence by the error sentence detection part;

a grammar correction part for generating a corrected sentence by performing language correction for grammar correction using the correction model on the sentence in which the spelling error is corrected by the spelling correction part; and

a post-processing part that performs post-processing of indicating a corrected portion during the language correction by the grammar correction part and outputs the corrected portion together with the corrected sentence.

97.The machine learning-based language correction system of claim 86, wherein the error sentence detection part classifies the sentence to be corrected into the error sentence and the non-error sentence according to reliability information recognized when the sentence to be corrected is classified.

~~108~~.The machine learning-based language correction system of claim ~~86~~, wherein the spelling correction part provides a spelling error occurrence probability value as reliability information when correcting a spelling error, the grammar correction part provides a probability value through an attention weight of language correction for the spelling error-corrected sentence as reliability information, and the post-processing part provides final reliability information of language correction for the sentence to be corrected by combining the reliability information provided by the spelling correction part and the reliability information provided by the grammar correction part.

~~119~~.The machine learning-based language correction system of claim ~~108~~, further comprising:

a language modeling unit that performs language modeling using a preset recommended sentence for the corrected sentence generated by the grammar correction part, between the grammar correction part and the post-processing part, wherein

the language modeling unit provides reliability information of the corrected sentence by combining a perplexity value and a mutual information (MI) value of a language model during the language modeling, and the post-processing part also combines the reliability information provided by the language modeling unit when providing the final reliability information.

~~12~~10.The machine learning-based language correction system of claim 1, further comprising:

a user dictionary including a source word registered by a user and a target word corresponding thereto, in which each of the source word and the target word is at least one word, wherein

the correction model learning part, when the word registered in the user dictionary is included in the data sets, replaces the word registered in the user dictionary and included in the data sets with a preset user dictionary marker to perform machine learning, and

the language correction unit, when the word included in the user dictionary is present in the sentence to be corrected, replaces the word included in the user dictionary and present in the sentence to be corrected with the user dictionary marker to perform language correction on the sentence to be corrected, and when the user dictionary marker is included in the corrected sentence, replaces the user dictionary marker included in the corrected sentence with the word registered in the user dictionary to correspond to a corresponding word in the sentence to be corrected.

~~13.A method for enabling a language correction system to learn a language correction model based on machine learning, the method comprising:~~

~~performing a learning processing including a supervised learning data labeling task, a machine learning data expansion task, and a parallel data~~

~~construction task for machine learning on a plurality of data sets including ungrammatical sentence data and error-free grammatical sentence data corresponding to the ungrammatical sentence data, respectively; and generating a corresponding correction model by performing supervised learning-based machine learning on the data sets on which the learning processing has been performed.~~

~~—— 14. The method of claim 13, wherein the machine learning data expansion task includes a data expansion task using letters formed from surrounding typing error characters around an in-position of a keyboard for typing letters included in the ungrammatical sentence data, and the parallel data construction task for machine learning includes a task of constructing parallel data using a parallel corpus formed by paring ungrammatical sentences unnecessary for correction and corresponding grammatical sentences.~~

~~—— 15. The method of claim 13, further comprising: performing pre-processing including a filtering task into a monolingual sentence and a data purification and normalization task by performing language detection on the data sets before performing the learning processing, wherein the performing of the pre-processing includes:~~

~~performing a translation on a large amount of ungrammatical sentence data in the data sets through the translation engine;~~
~~marking words, which are unregistered in a dictionary used by the translation engine, by using a preset marker, completing the translation on the large amount of ungrammatical sentence data, and then extracting the words marked by the preset marker; and~~
~~collectively correcting the extracted words into error-free words.~~

———16.The method of claim 15, wherein the collectively correcting of the words includes:

~~extracting the words marked by the preset marker;~~
~~checking a frequency of the extracted words;~~
~~arranging the words marked by the preset marker based on the checked frequency; and~~
~~collectively correcting the arranged words into error-free words.~~

———17.The method of claim 13, wherein the language correction system further includes a user dictionary including a source word registered by a user and a target word corresponding thereto, in which each of the source word and the target word is at least one word, and
the generating of the correction model includes replacing, when the word registered in the user dictionary is included in the data sets, the word registered in the user dictionary and included in the data sets with a preset

~~user dictionary marker to perform machine learning, thereby generating the correction model.~~

~~18.A method for enabling a language correction system to perform a language correction based on machine learning, the method comprising: performing spelling error correction on sentences to be corrected; and generating a corrected sentence by performing grammar correction by using a correction model on the spelling error-corrected sentence, wherein the correction model is generated by performing supervised learning-based machine learning on a plurality of data sets consisting of ungrammatical sentence data and error-free grammatical sentence data corresponding to the ungrammatical sentence data, respectively.~~

~~19.The method of claim 18, further comprising: before the performing the spelling error correction, performing a sentence segmentation, for the sentences to be corrected, in a unit of sentence and performing pre-process of tokenizing the segmented sentences; and classifying the sentences to be corrected that has been pre-processed into error sentences and non-error sentences by using a binary classifier, wherein~~

~~the classifying of the error sentences and the non-error sentences includes performing the spelling error correction when the sentence to be corrected is classified as the error sentence.~~

~~—— 20. The method of claim 19, wherein the classifying of the error sentences and the non-error sentences further includes classifying the error sentence and the non-error sentence according to reliability information recognized when the sentence to be corrected is classified.~~

~~—— 21. The method of claim 18, further comprising: after the generating of the corrected sentence, performing language modeling on the corrected sentence by using a preset recommendation sentence; and performing post-processing of indicating a corrected portion during the generating of the corrected sentence to output the corrected portion together with the corrected sentence.~~

~~—— 22. The method of claim 18, wherein the language correction system includes a user dictionary including a source word registered by a user and a target word corresponding thereto, in which each of the source word and the target word is at least one word, and wherein the method further comprises: before the performing the spelling error correction,~~

~~determining whether the word included in the user dictionary is included in the sentence to be corrected; and~~

~~replacing a word commonly included in the user dictionary and the sentence to be corrected with a preset user dictionary marker when the word included in the user dictionary is included in the sentence to be corrected, and~~

~~the method further comprises: after the generating of the corrected sentence, checking whether the user dictionary marker is included in the generated corrected sentence; and~~

~~generating a final corrected sentence, when the user dictionary marker is included in the generated corrected sentence, by replacing the word in the user dictionary corresponding to the word in the sentence to be corrected corresponding to a position of the included user dictionary marker.~~