Technology Reports

Automatic Scoring Deep Learning Language Processing

Technology to Grade and Correct Compositions in English

Service Innovation Department Hosei Matsuoka Toshimitsu Nakamura
Soichiro Murakami Atsuki Sawayama

English education in Japan in recent years emphasizes higher thinking and expression capabilities and requires well-balanced acquisition of the four skills of "listening," "reading," "writing" and "speaking." Among these, NTT DOCOMO has focused on the work of grading English compositions, i.e., "writing," and has developed technology to automatically grade and correct answers with AI. Once used to only evaluate the correctness of vocabulary and grammar, automatic grading can now also capture the meaning of an entire sentence to grade and correct it. This article describes an overview of this technology and its application.

1. Introduction

With globalization, English education in Japan in recent years has come to require well-balanced acquisition of the four skills of "listening," "reading," "writing" and "speaking," with an emphasis on higher thinking and expression capabilities. This has led to attention being focused on automatic grading systems that reduce the burden of grading and correcting as the number of English composition questions inevitably increases in various exams measuring English ability. Against this background, NTT DOCOMO has developed an English composition grading and correction technology that uses deep learning*1 to grade and correct learners' English compositions. This English composition grading/correction technology can grade English composition for Japanese sentences that don't exist in learning data by making computers learn large amounts of bilingual corpus*2 of Japanese to English

*1 Deep learning: A type of machine learning method that entails using large amounts of data to teach a computer capabilities that simulate human intelligence.

*2 Corpus: A language resource consisting of a large volume of text and utterances, etc. collected and stored in a database.

translations. The grading process entails expressing the meaning of the Japanese examination sentence as a vector, and generating an answer for it in English from that vector. Depending on the examination sentence, there could be several possible patterns of English that could be answers. The process of generating answers in English from the above Japanese sentence vectors aims to generate and grade English sentences that are closest to the expression of the learner's English composition to grade it. Therefore, if the learner's expression is different from the sample answer but nonetheless conveys the correct meaning, it will be graded highly.

In the past, an English composition was graded based on how close it was to a sample answer, i.e., how many words the answer contained with the same vocabulary and syntax (grammar) as the sample answer. However, with this technology it's also possible to grade for paraphrasing.

This article describes an overview of technology to grade and correct compositions in English.

2. Conventional Automatic Grading Technology

In the past, research on automatic grading of English compositions has generally involved technologies that grade based on how close the composition is to a sample answer at the grammar and vocabulary level. In 1966, Project Essay Grade (PEG®*3) [1], the first automatic grading system, was developed. Currently, various automatic grading systems such as Criterion®*4 [2] developed by Educational Testing Service (ETS) exist. These systems use the aforementioned technology.

English compositions can be broadly classified into two types: Japanese to English translation, and free English composition. Japanese to English translation entails replacing Japanese with English, while free English composition entails freely expressing one's ideas about a theme in English. While the above automatic grading technologies can be applied to both Japanese to English translations and free English compositions, what is actually seen as evaluation indicators are mostly such things as differences in grammar or vocabulary and length of sentence, etc. that are used to estimate the evaluation from sentence forms, but which are problematic when judging by correctly understanding the meaning of the overall sentence. English learners in the early stages of learning more often study Japanese to English translations than free English composition. For this reason, the grading and correction technology described in this article focuses on Japanese to English translations and focuses on indicators of whether the meaning of the examination sentence has been expressed.

3. Proposed Technology to Grade and Correct Compositions in English

As shown in **Figure 1**, this technology firstly entails preparing a large amount of data pairs (a Japanese-English bilingual corpus) consisting of Japanese sentences and sample answers for the Japanese to English translation problems. Then, deep learning is used to understand the meaning of entire sentences and construct a model that can be used for grading English translation answers. The bilingual corpus contains millions of translation pairs including spoken and written words to handle

*3 PEG®: A trademark or registered trademark of Measurement Incorporation.
*4 Criterion®: A trademark or registered trademark of Educational Testing Service.

a variety of sentences. This model is used to understand the meaning of the examination sentence and grade and correct the answer.

For learning, we used a recurrent neural network*5 encoder/decoder model*6, and input sequences of words obtained by dividing the words in the Japanese text for examination sentences into the encoder, and input sequences of words obtained by dividing the words in the English text of sample answers into the decoder.

Figure 2 shows the grading method using a model that has done this learning. For example, for the Japanese examination sentence "明日は晴れです," firstly the vocabulary in the sentence is

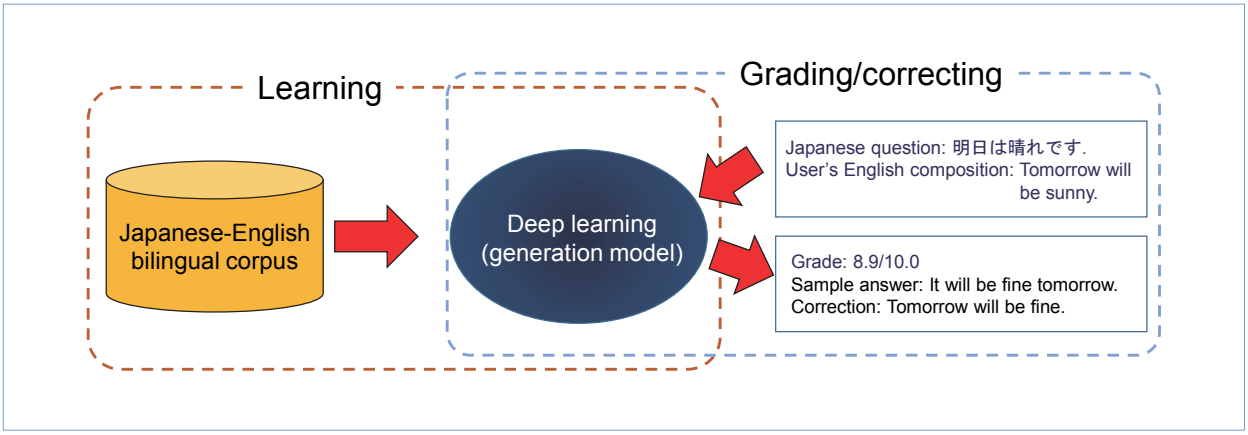

Figure 1 Application of deep learning

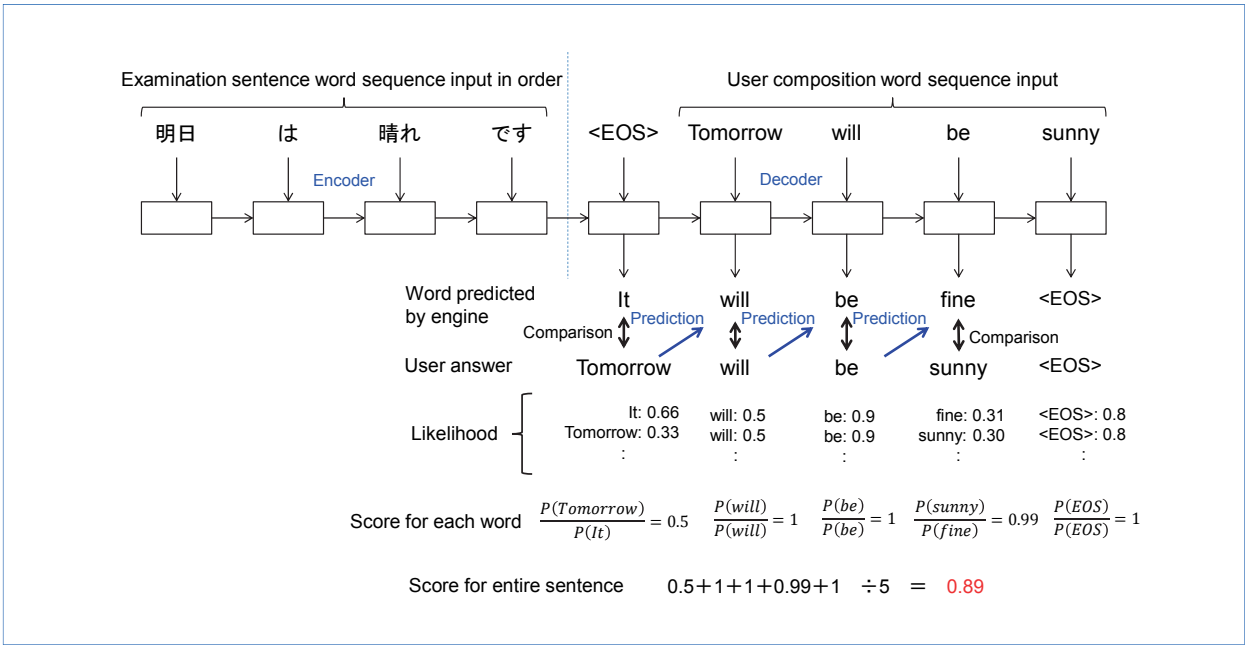

Figure 2 Grading method

*5 Recurrent neural network: A method of neural networking that entails a recurrent network structure in which the output of the intermediate layer is the input of the next step in a time series.

*6 Encoder/decoder model: A recurrent neural network structure that generates time series data from some time series data input.

divided into “明日,” “は,” “晴れ” and “です,” and then input in order into the encoder of the encoder/decoder model. Following, an End Of Sentence (EOS) symbol to indicate the end of the sentence is first input into the decoder section. Then, English translation of the examination sentences, and the first word of the predicted English text is output from the decoder. In the decoder, grading is performed by comparing each word of the user’s English composition input into the decoder with the predicted word. The decoder also outputs the next predicted word.

At first, triggered by <EOS> input, “It” is output, the first word of “It will be fine tomorrow,” which is the English translation of “明日は晴れです.” Here, if the user’s answer is “Tomorrow will be sunny,” the probability that the first word “Tomorrow” is output from the decoder first is calculated. If the likelihood[*7] of the first word output from the decoder is 0.66 for “It” and 0.33 for “Tomorrow,” the likelihood of “Tomorrow” divided by the highest likelihood of “It” is 0.5, which is the score awarded for the word.

Next, the first input of the user’s answer, “Tomorrow,” is processed as the next decoder input. The decoder fixes the first word to “Tomorrow” and predicts the next word, and “will” is the word with the highest likelihood. Since this matches the user’s answer, “1” obtained by dividing the likelihood of “will” by the likelihood of “will” becomes the score for this word. The calculations are performed in the same way up to the last word, and the average of the scores for all the words is used for the resulting grade of the entire sentence.

Grading is thus done by inputting the user’s answer sentence one word at a time into the decoder, predicting the next best word using the user’s answer, and comparing that word with the next word of the user’s answer. Accordingly, even if the user’s answer is expressed differently from the sample answer, the decoder creates a continuation of the English sentence based on the expression used by the user and compares that English sentence with the user’s answer, which enables handling of diverse expressions.

Moreover, if the user wasn’t able to complete a sentence, the decoder can write the rest of it. Correction is enabled by replacing erroneous words in the user’s answer with words with the highest likelihood and proposing subsequent text to improve the user’s answer.

Deep learning and a large-scale bilingual corpus make it possible to evaluate English sentences while understanding the meaning of Japanese examination sentences. The more data for pairs of examination sentences and answers, the more accurate grading and correction can be performed. Although there may be multiple answers depending on the question, bilingual pairs are created for each answer example, and making the English translation model learn with deep learning makes it possible to respond to a variety of answer examples. It’s also possible to grade and correct examination sentences that don’t exist in learning data, which we believe will be useful for learners’ selfstudy.

4. Grading Indicators and Grading Examples

In the English composition grading of this technology, the scores of 0 to 1 awarded by the decoder are multiplied by 10 to give a grade out of 10

*7 Likelihood: A numeric value that expresses the probability of guessing some result.

points. **Table 1** shows rough indicators actually created from scores for some questions and example answers.

Following is example of grading using this technology. The Japanese text of the examination sentence and the English text of the sample answer are as follows.

Examination sentence: このバスに乗れば，駅に着きます。

Sample answer: If you take this bus, you will get to the station.

Figure 3 shows this sentence with various example answers and their scores.

Sentence (1) "If you take this bus, you will get to the station." is exactly the same as the sample answer, and is awarded 10 points, and evaluated at level 5. Sentence (2) "This bus will take you to

Table 1　Grading indicators

5-grade evaluation level	Description	Score from grading
5	Meaning correctly conveyed. A fluent sentence in the native perspective.	9.0～10.0
4	Meaning correctly conveyed. Some improvements could be made from the native perspective.	8.0～9.0
3	Meaning mostly correctly conveyed. Improvements required.	7.0～8.0
2	Meaning may not correctly be conveyed. Contains a lot of unnatural language.	5.0～7.0
1	Meaning not correctly conveyed. Not a proper sentence.	0.0～5.0

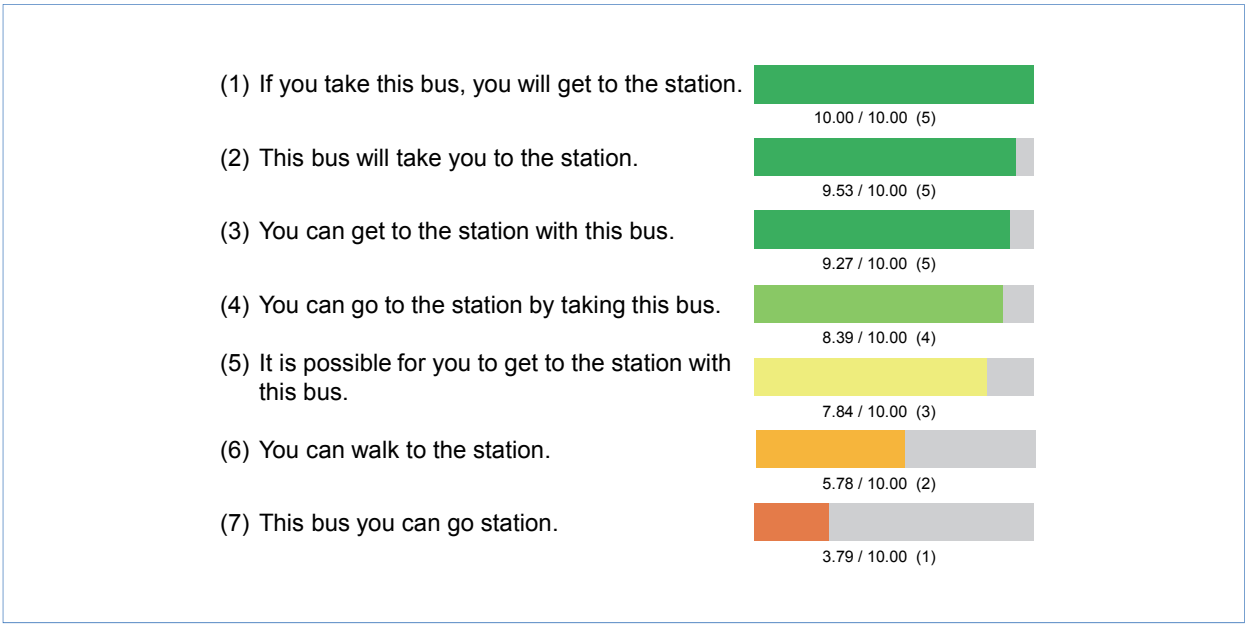

Figure 3　Example of grading

the station." has the bus as the subject, but correctly conveys the meaning, and is awarded 9.53 points and evaluated highly at level 5. In this way, a high grade can be awarded if the meaning is the same, but the sentence structure has been completely changed. Sentence (3) "You can get to the station with this bus." is also in a different form but the meaning is the same and so it is awarded 9.27 points with a high evaluation of level 5, while sentence (4) "You can get to the station by taking this bus." is not incorrect, however, the construction "by taking this bus" is not fluent in the native perspective. Therefore, the sentence is awarded 8.39 points and an evaluation of level 4. Similarly, sentence (5) is awarded 7.84 points and evaluation level 3 for its wordy construction. Sentence (6) "You can walk to the station." uses "walk" instead of "bus", and is therefore wrong, and is awarded 5.78 points and an evaluation level 2. Sentence (7) "This bus you can go station" is just a random attempt at stringing together the words used in the sample answer, and is not a proper sentence and thus is evaluated at level 1. In this way, even if the same words as the sample answer are used, if the meaning is not conveyed the sentence is given the low evaluation. Grading is thus done by understanding the structure and meaning of sentences.

In conventional English education, we hear that these Japanese to English translations are often marked as either correct or incorrect depending on whether they match the sample answer. However, this technology enables evaluation by enumerating fluency and conveyance of meaning, which we believe is an effective method of measuring the extent that meaning has been conveyed even if there are some errors. We also believe this system will be very useful for learners' self-study because it can provide corrections and feedback for creating better answers based on the user's answers.

5. Conclusion

This article describes development of a grading and correction technology that focuses on the meaning of Japanese to English translations. We believe that AI grading and corrections are useful in practicing the large number of Japanese to English translations required to master English composition. Going forward, we also intend to develop technology to handle free English compositions such as essays, etc. for tests such as The Eiken Test in Practical English Proficiency and the Test of English as a Foreign Language (TOEFL).

REFERENCES

[1] E. B. Page: "The Imminence of Grading Essays by Computer," Phi Delta Kappan, Vol.47, No.5, pp.238-243, 1966.
[2] J. Burstein, M. Chodorow and C. Leacock: "Criterion℠ Online Essay Evaluation: An Application for Automated Evaluation of Student Essays," IAAI, pp.3-10, 2003.