Interpreting Undeciphered Poems of Man'yoshu

2230007 Hiroaki Sasaki

Man’yoshu, Japan's oldest anthology of poetry, encompasses 4,516 poems, though 42 of them remain undeciphered. Previous studies assumed these compositions, written in Man’yogana (a writing system of using Chinese characters to represent Japanese phonemes), should be interpreted as Japanese. However, Vovin successfully demonstrated that Man’yoshu's poem number 9, which was previously undeciphered, became intelligible when it was read as Old Korean. The goal of this study is to decode the other undeciphered poems of Man'yoshu as Old Korean. We implement Vovin's method as a computational system to automatically or semi-automatically decipher undeciphered poems.

Our proposed method primarily involves two modules. The first module converts sequences of Chinese characters into sequences of phonemes using the “Chinese character phoneme dictionary.” The module then performs morphological analysis on these phoneme sequences using a Korean word dictionary, and finally generates sequences of Korean words. The second module enlarges our Chinese character phoneme dictionary by adding Middle Korean (MK) pronunciations. This is achieved by estimating Middle Korean pronunciations from Chinese pronunciations.

To implement the first module, we begin by creating the Chinese character phoneme dictionary which defines the phonetic symbols associated with each Chinese character. This dictionary compiles five types of pronunciations: Man’yogana, Idu (a writing system of using Chinese character to represent Korean phonemes), Middle Korean pronunciations derived from Late Han Chinese (LHC) pronunciations, Middle Korean pronunciations derived from Early Middle Chinese (EMC) pronunciations, and jeongyong pronunciations that represent Korean phonemes intended to convey the Chinese character’s meaning.

For a given input sequence of Chinese characters, the system converts it to multiple possible sequences of phonemes by looking up the Chinese character phoneme dictionary for each of the input Chinese characters. Next, for each phonetic sequence, morphological analysis is performed to obtain sequences of Middle Korean words. MeCab is employed for this process, since it is the only morphological analyzer for which a Middle Korean word dictionary is available. Specifically, this study utilizes the MkHanDic dictionary with 9,653 words as the MK word dictionary. Then, the most appropriate word sequences are chosen from all generated sequences using the following two-step procedures. (1) Any

grammatically incorrect word sequences are removed by the grammatical check. Specifically, we eliminate word sequences that begin with a verbal ending, begin with a specifier, begin with a dependent noun, end with a numeral, contain consecutive word endings, or include unknown words. (2) From the remaining valid word sequences, those consisting of the fewest words are chosen according to the minimum number count principle, which is a heuristic rule employed in morphological analysis to determine the optimal results. Finally, we manually translate the chosen Middle Korean word sequence, which may represent a possible interpretation of the poem, into Japanese, seeking to decipher the original undeciphered poem.

The second module addresses limitations in our available data. The materials we use to compile our Chinese character phoneme dictionary do not provide Middle Korean pronunciations derived from LHC and EMC for all Chinese characters. Therefore, models to predict Middle Korean pronunciations from Chinese pronunciations are trained. Specifically, we represent phonemes as sequences of International Phonetic Alphabet (IPA) symbols (phonemes) and train a sequence-to-sequence model to convert Chinese phoneme sequences into Middle Korean phoneme sequences. Each IPA symbol is represented as either a feature vector reflecting their phonetic features (called IPA feature vector), or a one-hot vector. Besides, a bidirectional or unidirectional Long Short-Term Memory (LSTM) is used as our sequence-to-sequence model. This results in proposing four models based on various combinations of IPA vector representations and bidirectional/unidirectional LSTMs. These models are trained using the set of Chinese characters associated with both the corresponding LHC or EMC phonemes and the MK phonemes.

Several experiments are carried out to evaluate our proposed method. First, our model to predict Middle Korean pronunciations from LHC or EMC pronunciations is evaluated. The accuracy of the model to convert LHC pronunciations to MK pronunciations is 0.839, while that of the model to convert EMC to MK is 0.800. Notably, the BiLSTM-IPA-IPA model, which is the BiLSTM model handling IPA feature vectors in both the encoder and the decoder, achieves the highest accuracy among our four proposed models, while also requiring the shortest training time.

Next, our proposed method is applied to poem number 9, which Vovin successfully deciphered, to verify its validity. Our method generates the same interpretation as Vovin, indicating the effectiveness of our method. This success demonstrates that our method has the potential to decipher other undeciphered poems besides poem number 9. Subsequently, our method is applied to six undeciphered poems to attempt their decipherment. As a result, we obtain the interpretation “Would you eat rice? Would you?”

for poem number 3889. The undeciphered section of poem number 3889 constitutes the final part of the poem. The preceding part of the poem is transcribed as “hitodama no sa aonaru kimi ga tada hitori mo aheri shi ameyo no,” which means “I met you, a wandering soul, blue, alone, on a rainy night.” The newly obtained interpretation for the undeciphered section is not clearly related to this preceding part, but it is not likely to be completely unrelated. Determining the validity of this interpretation remains a complex challenge.

Moreover, Man’yoshu also contains approximately 500 words of Makurakotoba, whose meanings are largely unknown. Makurakotoba is considered to be a rhetorical word that always modifies a specific word as its head; however, it is generally recognized that its origin remains unclear and that it has no semantic meaning. We consider that, like the undeciphered poems, Makurakotoba might also be represented by Korean phonemes. Therefore, we apply our proposed method to decipher six words of Makurakotoba. As a result, we obtain the interpretation of “father” or “paternal” for the Makurakotoba word “ashihikino.” This interpretation aligns reasonably well with its head “mountain.” However, since “ashihikino” has 31 written forms of Chinese characters and this feasible interpretation is obtained only from one of them, it is still uncertain that this Makurakotoba has been successfully deciphered.

The future work of this study includes a revision of the input data, an extension of materials for the Chinese character phoneme dictionary, and comprehensive evaluation. In this study, the text based on the Nishihonganji version is used as the input data. However, other Chinese characters from variant versions can also be used. For the Chinese character phoneme dictionary, it is promising to add more Chinese character pronunciations from additional texts, as well as to incorporate the Go-on pronunciations which were prevalent in Japan during the compilation of Man’yoshu. For the evaluation of the method, our method has been applied for only 6 of the 42 undeciphered poems and 6 of the 500 words of Makurakotoba. More comprehensive experiments are required to precisely evaluate the effectiveness of our proposed method. Moreover, augmentation of the Middle Korean dictionary MkHanDic, refinement of the method for selecting appropriate word sequences, and establishment of a methodology to assess the validity of the obtained interpretation of undeciphered poems are other important lines for the future direction.