1 Application Detail

1. The type of presentation (poster, short paper, long paper or panel)

   POSTER

2. A title

   Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations

3. A list of keywords (up to five)

   corpus linguistics, word correspondences, classical Japanese poetry, modern translations, asymptotic algorithm

4. The name, status and affiliation of the presenter (s)

   - Hilofumi Yamamoto, Tokyo Institute of Technology / University of California, San Diego
   - Bor Hodošček, Meiji University
   - Hajime Murai, Tokyo Institute of Technology

5. A contact email address

   - yamagen@ryu.titech.ac.jp

6. A postal address

   - Tokyo Institute of Technology: W1-8, 2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8550, Japan

2 A biography

Hilofumi Yamamoto is an associate professor at the Tokyo Institute of Technology. He earned a Ph. D. in Linguistics at Australian National University and is currently working on the mathematical modeling of vocabulary, linguistic change, and language complexity.

Bor Hodošček is a JSPS Postdoctoral Foreign Fellow at Meiji University. He earned a Ph. D. in Engineering from the Tokyo Institute of Technology and is currently working on the quantitative modeling of register in Japanese as well as exploring its role in writing assistance systems. His interests include quantitative linguistics, natural language processing, and educational technology.

Hajime Murai is an assistant professor at the Graduate School of Decision Science and Technology, Tokyo Institute of Technology. He earned a Ph. D. in Engineering from the Tokyo Institute of Technology. His majors are natural language processing, bibliometrics, and bible studies. He is currently working on the scientific analysis of text interpretation, and quantitative analysis within the humanities.

3 Abstract

Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations

Objectives: This project will develop an automatic word concordance system for parallel texts comprising of Classical Japanese poem texts and their associated modern translations. By using these parallel texts, we will clarify the details of language change within Japanese in an objective procedural manner that is not influenced by human observations.

Problem: Many scholars of classical Japanese poetry have tried to explain the constructions of poetic vocabulary using their intuition or the experiences they accumulated during their studies. Thus, they often produce modern Japanese translations through means of holistic explanations of each poem since they, even as specialist in classical Japanese poetry, cannot adequately explain the precise meanings of all words. Even if they could clearly explain all the words, their translations could include contradictions with their own explanations. This shows that translations include possibilities of non-literal elements, which are not expressible using ordinary word explanations.

To find the non-literal elements, we cannot manually match classical words in poem texts with modern words in translation texts. The problem of manual matchings of word for word from a parallel corpus, especially one comprised of classical texts, is that the act of judgment in identifying correspondences can lead to a loss of the original meaning of a word, since our present knowledge of classical words is conjectured and classified based on our knowledge of modern language.

To this end, it is necessary to employ computer assisted correspondence methods without relying on this human knowledge. We therefore use the asymptotic correspondence vocabulary presumption method(Murai, 2012) to estimate corresponding pairs of classical Japanese words and their modern Japanese translations.

Methods: Using the asymptotic correspondence vocabulary presumption method to classic texts and those modern translations, the proposed method allows the extraction of corresponding vocabulary pairs. A word in a poem text will be paired with every word in the corresponding translation text of the poem. This process is repeated for all of the words in the poem using our program. Our system will generate poem-word and translation-word patterns from Kokinshū poem #1 to #1000 respectively. Based on the frequency data of poem-word and translation-word patterns, we calculate mutual information (MI) scores for each pair. Then, for each iteration, we determine the best-scoring corresponding pair. After determining the best-scoring pair words, we remove all occurrences of it from both poem texts and translation, recalculate the MI scores with the remaining pairs, and finally determine the second best-scoring pair. This process is repeated until the MI score goes below a certain preset threshold value.

Materials: We will use the Kokinshū with ten corresponding sets of modern translations. The Kokinshū is an anthology compiled under imperial or-

ders (ca. 905). The Kokinshū consists of 1,111 poems including long poems (chōka) and head-repeating poems (sedōka) which are not short poems (tanka; 5/7/5/7/7 syllable style). We will only use the short poem form, which amounts to 1,000 poems, for stylistic consistency. The ten sets of modern translations of the Kokinshū are translated from 1927 to 1998 by ten Japanese poetry scholars.

This project has already begun: the parallel corpus of the Kokinshū has been constructed. We are now working on the development of computer software and the optimization of the calculation methods. As a result of our development and experimentation, we expect that literal words and non-literal elements will be extracted using our system.

References

Murai, Hajime (2012) "Semantic Networks between Texts in Different Language Based on the Asymptotic Correspondence Vocabulary Presumption Method", *The Computers and the Humanities Symposium*, Vol. 2012, No. 7, pp. 61–68, Nov.