Modality Transformation based on the Correspondence of Multimodal Information

HAMAZONO Yumi

People understand things more deeply through natural language.
In recent years, the development of computers has led to the development of methods for handling natural language, images and numerical information, and their capabilities have improved. While various things can now be handled on computers, each thing has been handled independently for each modality, and the relationship between modalities has not yet been captured. Therefore, this paper addresses the issue of capturing and generating correspondence between different modalities, with centered on language.

The first task is generating non-verbal sequences from language, the task of generating robot behaviours from natural language. As a method for handling language on a computer, the meaning of sentences are captured by using the frequency information of words in the sentence and the similarity of words in the surrounding words to create a distributed semantic representation that automatically assigns a meaning. Distributed representation is used to automatically assign meanings based on the frequency information of words in a sentence and the similarity of words by surrounding words. As a non-verbal series, this task deals with robot movements. First, in order to generate movements for the robot that are different from the movements of human joints, we propose a framework that can imitate human movements by combining basic movements that can be performed by a robot. By combining these basic movements, it is possible to express and generate complex movements. After achieving these goals, the task of learning the correspondence between words and actions is tackled. As a method for capturing the correspondence between the two, we use neural networks in this task. In particular, we propose and validate a neural network structure that can learn the correspondence between various robot actions and ambiguous expressions.

Secondly, as a task to generate language from non-verbal language, we tackle the task of generating live video in the general domain. As there is no dataset available for this task, the dataset is constructed by collecting the actual situation using an existing video corpus. The task is then divided into two subtasks, timing estimation and speech generation, to solve the task of generating the actual situation from the video alone. This is because the actual situation is text that is output along with the video, so it is necessary to control the timing of the start of the actual situation and the length of the actual situation text. For timing estimation, a framework that can estimate human behaviour is used. For speech generation, we use neural network models suitable for language generation, in particular Transformer.

Finally, we address the problem of generating language from non-verbal data, which arises when

using real world data and text, analyses the causes of the problem and propose a solution. In particular, we address the problem of generating summary text from time-series numerical data. Existing datasets for generating summary text for time-series numerical data include. The problem of inconsistent reference times inherent to data-to-text for time-series data, and the problem of being asked to produce output containing attributes that cannot be predicted from the input data, which can occur when using real-world data and text. By understanding the procedures for creating datasets and the characteristics of input and output data, we propose a way to solve these problems without significantly changing the existing framework.

Through these three efforts, we propose a method for changing the direction of modality with a focus on natural language, and discuss the future direction of the method.