Analysing the quality of storytelling using multimodal information

2210003 Yuta Ashikaga

Stories in human culture have established a long tradition. People tell stories for a wide variety of reasons, from simply entertaining, to transmitting knowledge across generations, to maintaining culture, to warning others of danger. This study focuses on storytelling, a communication method of telling stories. Storytelling is an interactive art that stimulates the listener's imagination and reveals the elements and images of a story through words and actions. If stories trace facts, stories are creative, deeply expressive of a person's inner life and important for understanding the human condition. The uses of storytelling are varied. For example, in product sales, telling the buyer the right story at the right time can convey that the product's value meets their needs, or interpreting the buyer's stated story can provide clues to finding a solution to a problem. According to the definition of storytelling, storytelling can also interact and actively create a reality in the mind, such as a vivid story, sensory images or events, based on the listener's own past experiences to the listener. It is then stated that a more complete story is unique and gives belief and understanding in the mind of the listener, making them a co-creator of the story. In other words, storytellers should work on improving their storytelling performance, as good or poor storytelling skills affect the images created for listeners and their understanding.

However, even if storytelling is done to someone, there needs to be someone to evaluate it, and it is not known whether the evaluation is valid. Furthermore, even after appropriate assessment, performance will not improve if people do not know how to improve their storytelling skills. Therefore, in this study, we present a framework in which the target speaker can be evaluated by a machine learning model using a storytelling dataset. By obtaining an appropriate and numerical evaluation of one's own storytelling skills, one can objectively know one's own skills. In addition, conventional research has used many modalities for estimating storytelling skills, such as prosody, gestures, and listeners' affusions, as well as the content of the speaker's utterances. It is not easy to know one's own skills, as a lot of equipment and manpower are indispensable to collect such data. Therefore, in this study, only data that is relatively easy to collect, such as the content of the speaker's speech, i.e. text data, was required, thus reducing the cost of data collection. However, some storytelling used in real life is based on personal experience. When storytelling such content, which varies greatly from person to person, it is difficult to evaluate all speakers appropriately and equally. For this reason, the content of what the speakers say is a description of a specific video, to reduce the variation between people.

It then helps speakers to improve their performance after receiving an appropriate evaluation by revealing not only the evaluation, but also the reasons that led to the evaluation. Specifically, using text features based on speech data from the speaker and image features based on image data from a specific video, machine learning was used to learn each of the seven storytelling skill items, and the results obtained were used to elucidate what good storytelling is.

In this study, data from storytelling to illustrate a particular video were used to experiment and discuss the estimation of speakers' storytelling skills and their storytelling prowess. There were seven storytelling skill assessment items, each scored on an eight-point scale, with 1 being the lowest rating, and the score for each item was set as the objective variable in the skill estimation. The features used as explanatory variables were created by combining three types of features: linguistic features taken from before one of the final layers of BERT, linguistic features from the output vector of CLIP's text encoder and image features from the output vector of CLIP's image encoder. A Gaussian kernel SVR was used for the model and the coefficient of determination R^2 was obtained by five-part cross-validation. Experimental results showed that BERT-only features had the highest accuracy in many evaluation items, but the highest accuracy was obtained when CLIP linguistic features were used in addition to BERT for items such as 'the scene was well described in words and the content of the story (information) was accurately conveyed'. In CLIP, even the text encoder is influenced by the image during the learning phase. From this, it was considered that the ability to visualise the scene when storytelling was evaluated. In the evaluation item "confidently explained", the CLIP image features were more accurate than the BERT-only features. The fact that CLIP's image features were effective in the item "confidence", which is apparently unrelated to images, was considered to be due to the fact that when speakers speak confidently, they have a clear image of the image in their mind.

The model used in this study is SVR, and time series are not considered when creating features. By using a model that can handle time series such as LSTM as a machine learning model, when speech is input in a time series, the scene estimated images are also arranged in a time series, and by comparing the speech and scene estimated images, it is possible to quantify how coherent the speech is and improve accuracy. In the future, in order to clarify storytelling skills, it will be necessary to create another set of features using CLIP, looking at the correlation with the evaluation items as well as the correspondence of the time-series data.