A Study on Building Training Datasets for AI/ML/NLP based Forecast Support Solution

Inkyung Kim[1], Heesun Park[1], Chanyun Yang[1], Hyesook Lee[1]
[1]AI Meteorological Research Division/NIMS

The Korea Meteorological Administration provides weather forecasters with an average of about 2TB of weather information per day through the operation of the system such as COMIS which supports various functions for forecasting tasks. However, there is a difficulty in not being able to utilize the whole. National Institute of Meteorological Sciences(NIMS) is promoting the development of a Meteorological Information Retrieval System based on AI(MRS-AI) using natural language processing for the function of searching data as part of the development of an Artificial Intelligence(AI) solution for forecasting support. MRS-AI is a search service that understands users' intentions and provides appropriate information when inputting natural language of voice and text. As the AI technology has the characteristics of dependent on data, it is necessary to build direct training datasets that meets the goal of improving the forecasting environment. In this study, the role and main functions of the MRS-AI system are introduced through COMIS analysis to apply the AI solution on Prediction Support. We developed the methodology of building training datasets for the search function based on natural language processing, and constructed the training datasets in four steps. The entire URL of the target system, COMIS, was converted into a pseudo URL, and the criteria were prepared, including deriving considerations when selecting a search keyword through the argument analysis of the pseudo URL. Keywords that are mapped for each COMIS page were extracted considering established standards. Training datasets related to natural language sentence for the language model was constructed by applying the method of the natural language generation in the template format. In addition, we reconstructed training datasets by applying grouping method that merges URLs with similar data properties. As a result of training the language model with the reconstructed training datasets, the performance was improved.

Key words: Forecast, AI Solution, AI Retrieval System, RPA

※ This work was funded by the Korea Meteorological Administration Research and Development Program “Assistant Technology and Its Application for Weather Forecasting Process”under Grant (KMA2021-00123).