AI startup Upstage has achieved world-class research results in the field of natural language processing.

Upstage (CEO Kim Sung-hoon, www.upstage.ai) announced that it has reaffirmed its global top AI technology by presenting two papers at EMNLP 2023, the world's most prestigious conference in the field of natural language processing.

EMNLP 2023 is an abbreviation for Empirical Methods in Natural Language Processing, and it is the world's most prestigious conference that deals with research related to natural language processing approaches based on language data, such as AI translation, chatbots, and machine reading.

Last year, EMNLP 2022 received a total of 3,242 papers, of which only 715 were passed, recording an adoption rate of 22%. EMNLP 2023 will be held in Singapore from December 6th to December 10th, and world-renowned AI companies such as Google, Apple, Amazon, and Baidu will participate.

The two papers adopted this time are NLP research results related to the Korean language, and they were conducted in collaboration with the research team of Professor Lim Hee-seok at Korea University, led by Upstage's Tech Lead Park Chan-joon.

The first paper, 'KEBAP: Korean Error Explainable Benchmark Dataset for ASR and Post-processing', is a paper that builds a new benchmark dataset related to Korean speech recognition post-processors, and proposes a new evaluation methodology to evaluate and identify the weaknesses of speech recognition models.

This paper points out the problems of traditional evaluation methods that do not provide accurate information on the weaknesses of speech recognition models by considering two aspects: speech and text levels, and proposes a research method that improves the explainability of the model by integrating

speech and text level errors.

It has subdivided 37 speech level types considering background noise and speaker characteristics, and 13 text level error types, and analyzed them by applying the proposed evaluation method to commercialized speech recognition systems such as Google Cloud Speech Recognition and CLOVA.

The second paper, 'CHEF in the Language Kitchen: A Generative Data Augmentation Leveraging Korean Morpheme Ingredients', is a paper that proposes a new data augmentation technique that takes advantage of the characteristics of the Korean language.

Unlike English, Korean is composed of small units called morphemes, and the meaning of the sentence changes depending on the combination of morphemes. For example, by combining the morphemes 'bap' and'mok', you can create various sentences such as 'eat rice', 'ate rice', and 'want to eat rice', but if you arbitrarily augment the data without considering the characteristics, there is a blind spot that the meaning of the sentence changes or that an unnatural sentence is generated.

The paper proposes a methodology that reflects the characteristics of Korean language to generate natural sentences even with the same ingredients, and a methodology that generative language models augment data by varying the combination of Korean morphemes.

Upstage's achievements at EMNLP 2023 are a series of accomplishments at global conferences. Upstage has achieved the most research results among domestic companies by presenting seven papers at ICML 2023-DMLR, the most prestigious workshop in the field of Data-Centric AI, in June.

In addition, Upstage has achieved the feat of publishing 100 AI papers at home and abroad and achieving paper adoption in all of the top seven NLP conferences based on the Google Scholar ranking in just three

years since its inception.

Google Scholar ranking is a prestigious indicator that measures the influence of conferences by evaluating the number of citations of papers. The top seven NLP conferences include ACL, EMNLP, NAACL, TACL, COLING, LREC, and WMT, and Upstage has achieved research results in all conferences except TACL, which is classified as a journal.

Upstage CEO Kim Sung-hoon said, "We are very pleased to be able to achieve research results at various global conferences, including this EMNLP 2023." "Upstage will do its best to make everyone use the highest performance AI more conveniently based on research results through continuous R&D investment."