Long Text Segmentation by String Vector based KNN

Taeho Jo

Department of Liberal Art, Hongik University, 2639 Sejongro, Sejong, South Korea

tjo018@hongik.ac.kr

***Abstract*—In this research, we propose the string vector based version of KNN as the approach to the text segmentation. The text segmentation may be interpreted into the text classification, and encoding texts into string vectors improved previously the text classification performance. In this research, we encode sentence pairs or paragraph pairs into string vectors, and apply the string vector based version of KNN to the classification task mapped from the text segmentation. As the benefits from this research, we expect the better performance, the more compact representation, and the more transparency by doing so. The goal of this research is to improve the performance of the text segmentation system.**

***Keyword*—Text Segmentation, String Vector based AHC, Semantic Operation**

Taeho Jo (M’97–AM’12) This author became a Member (M) of IEEE in 1997, and an Associate Member (AM) in 2012. He was born in 1970, South Korea. He received his Bachelor degree from Korea University in 1994, his Master degree from Pohang University of Science and Technology in 1997, and his PhD degree from University of Ottawa in 2006. His research area spans mainly over text mining, neural networks, machine learning, and information retrieval. He has the four year experience of working for industrial organizations and ten year experience of working for academic ones. Now, he has published more than 100 research papers in peered journals and proceedings, and he is working for Hongik University as a faculty member.