

K-means clustering to improve the accuracy of decision tree response classification.

ABSTRACT

The use of deep generation with statistical-based surface generation merits from response utterances readily available from corpus. Representation and quality of the instance data are the foremost factors that affect classification accuracy of the statistical-based method. Thus, in classification task, any irrelevant or unreliable tagging of response classes represented will result in low accuracy. This study focused on improving dialogue act classification of a user utterance into a response class by clustering the semantic and pragmatic features extracted from each user utterance. A Decision tree approach is used to classify 64 mixed-initiative, transaction dialogue corpus in theater domain. The experiment shows that by using clustering technique in pre-processing stage for re-tagging response classes, the Decision tree is able to achieve 97.5% recognition accuracy in classification, better than the 81.95% recognition accuracy when using Decision tree alone.

Keyword: Classification; Clustering; K-means; Decision tree; Natural language generation; Dialogue systems.