World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:10, No:01, 2016

Improving Topic Quality of Scripts by Using Scene Similarity Based Word Co-Occurrence

Authors : Yunseok Noh, Chang-Uk Kwak, Sun-Joong Kim, Seong-Bae Park

Abstract : Scripts are one of the basic text resources to understand broadcasting contents. Since broadcast media wields lots of influence over the public, tools for understanding broadcasting contents are more required. Topic modeling is the method to get the summary of the broadcasting contents from its scripts. Generally, scripts represent contents descriptively with directions and speeches. Scripts also provide scene segments that can be seen as semantic units. Therefore, a script can be topic modeled by treating a scene segment as a document. Because scripts consist of speeches mainly, however, relatively small co-occurrences among words in the scene segments are observed. This causes inevitably the bad quality of topics based on statistical learning method. To tackle this problem, we propose a method of learning with additional word co-occurrence information obtained using scene similarities. The main idea of improving topic quality is that the information that two or more texts are topically related can be useful to learn high quality of topics. In addition, by using high quality of topics, we can get information more accurate whether two texts are related or not. In this paper, we regard two scene segments are related if their topical similarity is high enough. We also consider that words are co-occurred if they are in topically related scene segments together. In the experiments, we showed the proposed method generates a higher quality of topics from Korean drama scripts than the baselines.

Keywords : broadcasting contents, scripts, text similarity, topic model
Conference Title : ICAIPR 2016 : International Conference on Artificial Intelligence and Pattern Recognition
Conference Location : Singapore, Singapore
Conference Dates : January 07-08, 2016