سومین همایش ملی
زبان شناسی رایانشی
۲۸ و ۲۹ آبان ۱۳۹۳
دانشگاه صنعتی شریف

Persian Metaphor Frequency Prediction through Latent Dirichlet Allocation Model

Abstract

In this paper, we aim to predict the approximate frequency of metaphorical concepts in Persian language. As a first step, we apply LDA topic modeling on so-called Bijankhan corpus to extract topics. The extracted topics carry the words which share the most natural semantic proximity. Then, we develop a system for classifying natural and metaphorical sentences. Using the words of the topics, our system determines an overall topic for each sentence in the corpus. This system works on the assumption that if the overall topic of the sentence diverges from the topic of a word in the sentence, metaphoricity is detected. We have evaluated the system manually on 100 sentences and achieved the f-measure of 68.17%. Finally, we experiment and conclude that every at least two and at most four sentence seen in the corpus carries metaphoricity.

Keywords: Frequency, Persian language, LDA, Metaphoricity