Network Word Discovery Framework Based on Sentence Semantic Vector Similarity

Authors : Ganfeng Yu, Yuefeng Ma, Shanliang Yang

Abstract : The word discovery is a key problem in text information retrieval technology. Methods in new word discovery tend to be closely related to words because they generally obtain new word results by analyzing words. With the popularity of social networks, individual netizens and online self-media have generated various network texts for the convenience of online life, including network words that are far from standard Chinese expression. How detect network words is one of the important goals in the field of text information retrieval today. In this paper, we integrate the word embedding model and clustering methods to propose a network word discovery framework based on sentence semantic similarity (S^3-NWD) to detect network words effectively from the corpus. This framework constructs sentence semantic vectors through a distributed representation model, uses the similarity of sentence semantic vectors to determine the semantic relationship between sentences, and finally realizes network word discovery by the meaning of semantic replacement between sentences. The experiment verifies that the framework not only completes the rapid discovery of network words but also realizes the standard word meaning of the discovery of network words, which reflects the effectiveness of our work.

Keywords : text information retrieval, natural language processing, new word discovery, information extraction
Conference Title : ICWAIM 2022 : International Conference on Web-Age Information Management
Conference Location : Kuala Lumpur, Malaysia
Conference Dates : August 30-31, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:08, 2022 publications.waset.org/abstracts/153917.pdf