World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:12, No:10, 2018

Weighted-Distance Sliding Windows and Cooccurrence Graphs for Supporting Entity-Relationship Discovery in Unstructured Text

Authors : Paolo Fantozzi, Luigi Laura, Umberto Nanni

Abstract : The problem of Entity relation discovery in structured data, a well covered topic in literature, consists in searching within unstructured sources (typically, text) in order to find connections among entities. These can be a whole dictionary, or a specific collection of named items. In many cases machine learning and/or text mining techniques are used for this goal. These approaches might be unfeasible in computationally challenging problems, such as processing massive data streams. A faster approach consists in collecting the cooccurrences of any two words (entities) in order to create a graph of relations - a cooccurrence graph. Indeed each cooccurrence highlights some grade of semantic correlation between the words because it is more common to have related words close each other than having them in the opposite sides of the text. Some authors have used sliding windows for such problem: they count all the occurrences within a sliding windows running over the whole text. In this paper we generalise such technique, coming up to a Weighted-Distance Sliding Window, where each occurrence of two named items within the window is accounted with a weight depending on the distance between items: a closer distance implies a stronger evidence of a relationship. We develop an experiment in order to support this intuition, by applying this technique to a data set consisting in the text of the Bible, split into verses.

Keywords : cooccurrence graph, entity relation graph, unstructured text, weighted distance
Conference Title : ICCLTM 2018 : International Conference on Computational Linguistics and Text Mining
Conference Location : New York, United States
Conference Dates : October 08-09, 2018