Alian M., Awajan A., Ramadan B. 2019. Unsupervised Learning Blocking Keys Technique for Indexing Arabic Entity Resolution. International Journal of Speech Technology, Vol. 22, Number 3. DOI 10.1007/s10772-018-9489-6 (ISI, Springer - Scopus).

Abstract

Attribute values in textual datasets are subjects of different types of errors due to the data entry processes such as typographical errors, pronunciation errors or dialects alterations. These errors make the entity resolution process more challenging. The iterative blocking indexing technique can be used for correcting this type of errors mainly in query access where the records are stored into more than one block. Blocking indexing technique selects a subset of object pairs saved in the same block for later detailed computation for similarity discarding other pairs in other blocks considering them as irrelevant. This work aims to solving such problems for Arabic texts. It proposes to adapt a specific model for learning blocking keys and analyze its performance for Arabic datasets. The resulted blocking keys are passed as blocking keys to Dynamic Aware Inverted Index (DySimII) that worked efficiently with Arabic datasets. The model is tested against a telephone book dataset that contains duplicates and errors in attribute values according to phonetic and typing errors. The results reach a matching accuracy of 84% for using learned keys with small number of corrupted attributes while the performance is declined with the increase of the number of corrupted attributes