Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 95 (2016) 153 – 158

Procedia Computer Science

Complex Adaptive Systems, Publication 6
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2016 - Los Angeles, CA

Entity Resolution Using Convolutional Neural Network

Ram Deepak Gottapu[a], Dr. Cihan Dagli[a], Dr. Bharami Ali[a]*

Missouri University of Science and Technology, Rolla, MO, 65409, USA

Abstract

Entity resolution is an important application in field of data cleaning. Standard approaches like deterministic methods and probabilistic methods are generally used for this purpose. Many new approaches using single layer perceptron, crowdsourcing etc. are developed to improve the efficiency and also to reduce the time of entity resolution. The approaches used for this purpose also depend on the type of dataset, labeled or unlabeled. This paper presents a new method for labeled data which uses single layered convolutional neural network to perform entity resolution. It also describes how crowdsourcing can be used with the output of the convolutional neural network to further improve the accuracy of the approach while minimizing the cost of crowdsourcing. The paper also discusses the data pre-processing steps used for training the convolutional neural network. Finally it describes the airplane sensor dataset which is used for demonstration of this approach and then shows the experimental results achieved using convolutional neural network.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: word stemming; word embedding; convolutional neural network; crowdsourcing, hybrid machine-human model

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: rgrk6@mst.edu

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2016.09.306

1. Introduction

Entity resolution or record linkage is the task of identifying records that refer to same entity in a dataset or across multiple datasets. It becomes especially important when merging data from different sources. For example if we consider the product descriptions posted on a certain e-commerce website and the company wanted to classify these records based on the product model using the product descriptions as shown below:

Table 1. Product Dataset

Record	Product Description	Product Label
r_1	iPhone 5 16-white	Iphone 2015
r_2	iPhone 5th generation 16GB WiFi White	Iphone 2015
r_3	Samsung Galaxy Tab E 9.6" with 16 GB and Wifi (Black)	Galaxy 2015
r_4	Samsung Galaxy Tab E 9.6 inches 16 -Black	Galaxy 2015
r_5	Apple iPhone 5 16GB WiFi White	Iphone 2015

From the table we can infer that records r1, r2 and r5 refer to one product label/entity and r3 & r4 refer to another product label/entity. Since our approach deals with labeled data, we can consider the product label/entities as labels. However, the descriptions are not completely identical and hence we cannot use simple string comparisons to identify which records are identical and link them to product label. Probabilistic method is the most common algorithm used for entity resolution, as set forth by Howard Newcombe et al.[1, 2] , and formalized by Fillegi and Sunter [3]. However, it is not effective for above example as the method requires dataset to have multiple columns of data to find records that have high probability of being similar [4]. [5] described a hybrid machine-human approach which can be applied to datasets having single columns of sentences as show in above table. It uses Jaccard similarity (machine part) and crowdsourcing (human part) to perform entity resolution. In hybrid machine-human model the machine part is usually an algorithm and the human part is usually a task performed by the humans. For cases such as entity resolution, the algorithm part performs certain amount of record links and those linked records are then uploaded as small tasks on the crowdsourcing platform. These tasks are solved by people who get paid for each task they complete. The tasks are usually referred to as HIT's (human intelligence tasks) and the people working on those tasks are called crowd [5].

This is a very efficient way to perform entity resolution but using crowd for such purpose costs money. If there are millions of records on which entity resolution has to be performed, then the obvious goal is to reduce the tasks given to the crowd. This can be achieved if the machine part successfully classifies majority of the dataset. Similarity functions such as Jaccard similarity are not efficient for such a task as it requires generating record pairs in order to find which records are identical. For example if we have n records then we have to generate $n(n-1)/2$ record pairs and perform comparisons using similarity function in order to find out which records have high similarity. In addition to that, if the records are too distinct, similarity functions will identify those records as dis-similar records.

For example, Jaccard similarity over two records is defined as the size of the set intersection divided by the size of the set union.

$J(r_5, r_2) = (Iphone, 16GB, white, WiFi)/(Iphone, 16GB, white, Apple, 5th, 5, generation, WiFi) = 0.5$

Similarly $J(r_1, r_2) = 0.25$ and $J(r_3, r_4) = 0.59$. The records whose similarity is above a certain threshold are converted into tasks and uploaded in the crowdsourcing platform [5]. The similarity value is usually chosen above 0.5 so that the records having more than 50% similarity are sent to the crowd. However, we can observe cases like $J(r_1, r_2)$ whose similarity is less than threshold and yet they belong to same product label. The reason is that the words used are completely different to describe the same product. Hence we need a model that can assign correct labels even when the words used to describe the product are different and also which can reduce the tasks that have to be given to the crowd. In this paper, we also use hybrid machine-human approach but the machine part uses