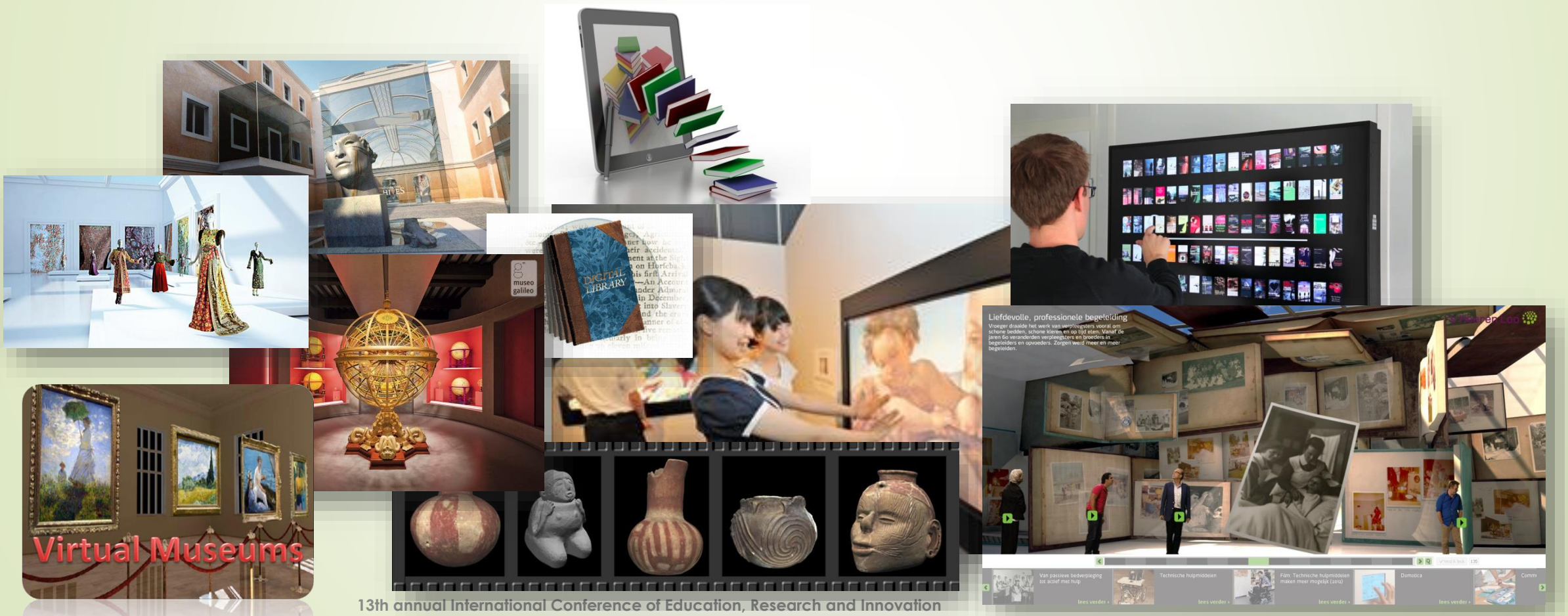
DATA DISCOVERY AND DISTRIBUTED REPRESENTATION FOR BETTER CULTURAL HERITAGE OBSERVATION AND LEARNING

Desislava Paneva-Marinova, Jordan Stoikov, Alexandra Nikolova, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

Lilia Pavlova, Laboratory of Telematics, Bulgarian Academy of Sciences

Introduction

This research work is focused on the content synthesizing activity, striving to deliver solutions for enhanced learning experience in the systems for digital cultural assets.



Main aim

- To propose a learning approach that support the educational process by facilitating a more effective learning content exploration, using automatic data discovery methods linking synonym learning concepts.
- The challenge for the human contribution:
  - the need of domain-specific experts with former knowledge on what values would most likely match in order to label matching entity pairs faster and more accurately, as derived from separate data pools with more than a few thousands of records.

Authomatic data discovery

- Data discovery is the collection and analysis of data from disparate data sources in order to gain additional insight of the data meaning from hidden patterns and trends.
- The typical process flow of similar record linkage consists of two functions:
  - (1) The match functions discover duplicated records, referring to the same entity, e.g. the same person;
  - (2) The merge function retrieves compound information about the matched entities from databases of different source that contain multiple records that represent the same person but contain different attributes.
- These methods and in particular the method of distributed representation could produce the required improvements to the learning experience of the users through facilitating the retrieval and analysis of specific data curation representations.

Use case

- The task: identifying tuples pairs that represent the same entity.
- For example, the tuple ⟨Alexander The Great, King of Macedonia⟩ and ⟨Alexander III the Great, Basileus of Macedonia⟩ refer to the same person as presented in the following table.

Tuple ID	A_1:Name	A_2:City
t_1	Alexander The Great	Pella
t_2	Alexander of Macedon	Pella

Our solution (1)

- we need to transform data to a vector representation in order to ration the resemblance between two tuples by estimating the distance between their adjacent vectors (next table).

Input	:	**Tuple t, a pre-trained dictionary such as GloVe**
Output	:	DR $v(t)$ for t
For	:	each attribute A_k of t executes the following:
Step 1	:	Tokenize $t[A_k]$ into a set of words W
Step 2	:	Look up vectors for tokens $w_l \in W$ in GloVe
Step 3	:	$V_k(t)$:= average of vectors of tokens in $t\ [A_k]$
Step 4	:	$v(t)$:= concatenation of $v(t[A_k])$, for $k \in [1, m]$

Our solution (2)

- If *T* is an entities set with *m* attributes and *n* tuples $\{A_1 \ldots\ldots A_m\}$. Those attributes could be derived from multiple tables. The value of attribute A_i from tuple *t* is denoted by $t[A_i]$. The issue with entity resolution to determine which tuples pair matches to the same entity if we have distinctive tuple pairs (t,t_i) from T where $t \neq t_i$.
- We alter the tuple into a vector representation, allowing to measure the resemblance between the two tuples, by calculating the distance between their respective vectors.
- If we have a tuple t with m attributes $\{A_1 \ldots\ldots A_m\}$, then values $t[A_k]$ is represented by vector v $t[A_k]$and tuple *t* is represented by vector v(t). Word x is represented by vector v(x) and $|v|$ is the number og dimensions number of vector *v*.
- The approach for computing v(t) is as follows: A standard tokenizer is used to break into individual word each attribute value $t[A_k]$. In this process, each separate piece of the broken up attribute is called a token. The pre-trained dictionary GloVe (a learning algorithm for obtaining vector representations for words) [7][8] is looked up for the d-dimensional v(x) vector corresponding to each token (word) (x). By averaging the x token's vectors in $t[A_k]$ is derived the vector representation for an attribute value $v(t[A_k])$. The concatenation of all vectors $v(t[A_k])(k \in [1;m])$ is the vector representation of v(t) of tuple t, as described in the Simple Averaging Algorithm.

Our solution (3)

- Based on the input from previous tables we will compute v(t) and the similarity between two vectors. From our sample data we obtain $v_1[t_1]$ = [0.45, 0.8, 0.85] and $v_1[t_2]$ = [0.4, 0.8, 0.75]. $v_2[t_1] = v_2[t_2]$ = [0.1, 0.1, 0.2]. Subsequently the Distributed representations for t1 and t2 are derived from the concatenation of the distributed representations of A_1 and A_2.

Word	DR (Distributed Representation)
Alexander	[0.5, 0.8, 0.8]
The Great	[0.4, 0.9, 0.9]
Of Macedon	[0.3, 0.8, 0.7]
Pella	[0.1, 0,1, 0.2]

This work is partly funded by the research projects:

Information and Communication Technologies for a Single Digital Market in Science, Education and Security *National Research Program* approved by DCM No 577 of 17 August 2018.

- Contractor: Bulgarian Academy of Sciences
- Duration: 36 months
- Funded by: Bulgarian Ministry of Education and Science

CLaDA-BG: the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH

- Contract № DO01-164/28.08.2018
- Contractor: Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
- Duration: 60 months
- Funded by: Bulgarian Ministry of Education and Science

Contact to the Authors:

Desislava Paneva-Marinova

Associate Professor, Ph. D.

Chair of the Mathematical Linguistics Department

Institute of Mathematics and Informatics

Bulgarian Academy of Sciences

8, G. Bonchev Str., 1113 Sofia, Bulgaria

Tel. +359 2 979 2874 (office)

e-mail: dessi@cc.bas.bg

Web site: http://mdl.cc.bas.bg/dessi

Jordan Stoikov

Ph.D. student

Mathematical Linguistics Department

Institute of Mathematics and Informatics

Bulgarian Academy of Sciences

8, G. Bonchev Str., 1113 Sofia, Bulgaria

e-mail: jstoikov@shieldui.com

Alexandra Nikolova

Ph. D. Student

Mathematical Linguistics Department

Institute of Mathematics and Informatics

Bulgarian Academy of Sciences

8, G. Bonchev Str., 1113 Sofia, Bulgaria

e-mail: alxnikolova@abv.bg

Lilia Pavlova

Assistant Professor, Ph. D.

Laboratory of Telematics

Bulgarian Academy of Sciences

8, G. Bonchev Str., 1113 Sofia, Bulgaria

e-mail: lilia.pavlova@gmail.com