IRCDL 2019, Pisa, Italy

Learning to Cite: Transfer Learning for Digital Archives

Dennis Dosso, Guido Setti and Gianmaria Silvello

Information Management Systems Research Group (IMS)
Department of Information Engineering

University of Padua

dosso@dei.unipd.it
silvello@dei.unipd.it

Scientific research relies more and more on data

**Strong demand to give databases the same
scholarly status of traditional references**

**Data needs to be retrieved, accessed and cited
as traditional scientific articles are**

**Automatic Creation of text snippets
for citing datasets
with a variable granularity**

LtC (Learning to Cite):

- **automatically** create a **model**
- **from data (EAD XML files)**
- **no manual intervention from data curators**
- **no modification of the data**

EAD Encoded Description Files

1. **XML** files used to describe the content of a whole archive
2. They are a means to describe the internal logic of an archive
3. They are **deep**
4. **Not easy to navigate** for a user
5. Wide variability in the use of **tags**, not straightforward to build a rule-based system
6. Every node is a potential **citation unit**

Thus, they are a good test-bed for a data-citation system

Collection $\mathcal{C}$ (XML files)

$$T = \{t_1, \dots, t_n\}$$

XML trees

$$\mathcal{H} = \{H_1, \dots, H_m\}$$

**Human-readable
citations**

$$\langle t_i, H_i \rangle$$

training data

**Learning to
Cite**

Citation Model

Validation and test data

$$\langle p_t, t_t \rangle$$

p_t XPath to a citable unit,

t_t XML tree containing the citable unit

machine-readable citation
generated from the system for e_k

$$MC_k = \{p_1, p_2, \dots, p_n\}$$

p_i are the paths composing
the citation

Ground Truth citation
for e_k

$$GTC_k = \{p_1, p_2, \dots, p_m\}$$

p_i are the paths composing
the citation

Optimization Measures for Validation and Test

$$precision = \frac{|MC_k \cap GTC_k|}{|MC_k|}$$

$$recall = \frac{|MC_k \cup GTC_k|}{|MC_k|}$$

$$f - score = 2 * \frac{precision * recall}{precision + recall}$$

Transfer Learning

COLLECTION 1: Library of Congress (LoC)

- 2083 XML files
- 25 EAD files were randomly selected
- 100 citation units extracted by hand
- Citations built by hand from each citation unit as ground truths for training and testing

-
- 5 different and heterogeneous sources:
 1. University of Chicago Library
 2. University of Maryland Libraries
 3. National Archives, Den Haag
 4. Syracuse University
 5. WorldCat aggregate collection
 - 10 citable units from each subset
 - 50 citable units in total

COLLECTION 2: EAD various

Experiment 1

Model generated from LoC and tested on LoC and on EAD various (transfer learning)

training set size	Precision LoC	Precision EAD	Recall LoC	Recall EAD	f-score LoC	f-score EAD
20	0.8819	0.5289	0.8232	0.4188	0.8441	0.4584
30	0.9034	0.5283	0.8089	0.4101	0.8465	0.4525
40	0.8748	0.5282	0.8312	0.4254	0.8447	0.4623
50	0.9239	0.5277	0.8045	0.4043	0.8531	0.4488
60	0.9333	0.5414	0.7723	0.3955	0.8366	0.4474
70	0.9012	0.5284	0.8106	0.4131	0.8462	0.4547
80	0.9152	0.5281	0.8055	0.4065	0.8506	0.4503

- With the transfer learning, performances are halved
- LtC performs best on the set it was trained
- High values even with small training set (regularity of the structure)

Experiment 2

Model generated from EAD various collection 6 times, leaving one sub-collection out each time. Test on LoC

collection left out	Precision LoC	Recall LoC	f-score LoC
none	0.4143	0.4531	0.4199
Chicago	0.4366	0.4498	0.4334
MdU	0.4577	0.4518	0.4439
NL-HaNA	0.4191	0.4530	0.4236
syracuse	0.4144	0.4430	0.4178
worldcat	0.4204	0.4485	0.4238

- Aim of the experiment is to find if one sub-collection is more informative than the others
- The differences between the collections are not statistically relevant

Conclusions

- **LtC behaves very well even with a training set of 20**
- **Performs at best when trained on the collection where it will applied**
- **The impact of the sub-collections is marginal**