ESSTER at the EYRE 2020 Entity Summarization Task

Qingxia Liu, Gong Cheng, and Yuzhong Qu

State Key Laboratory for Novel Software Technology, Nanjing University, China

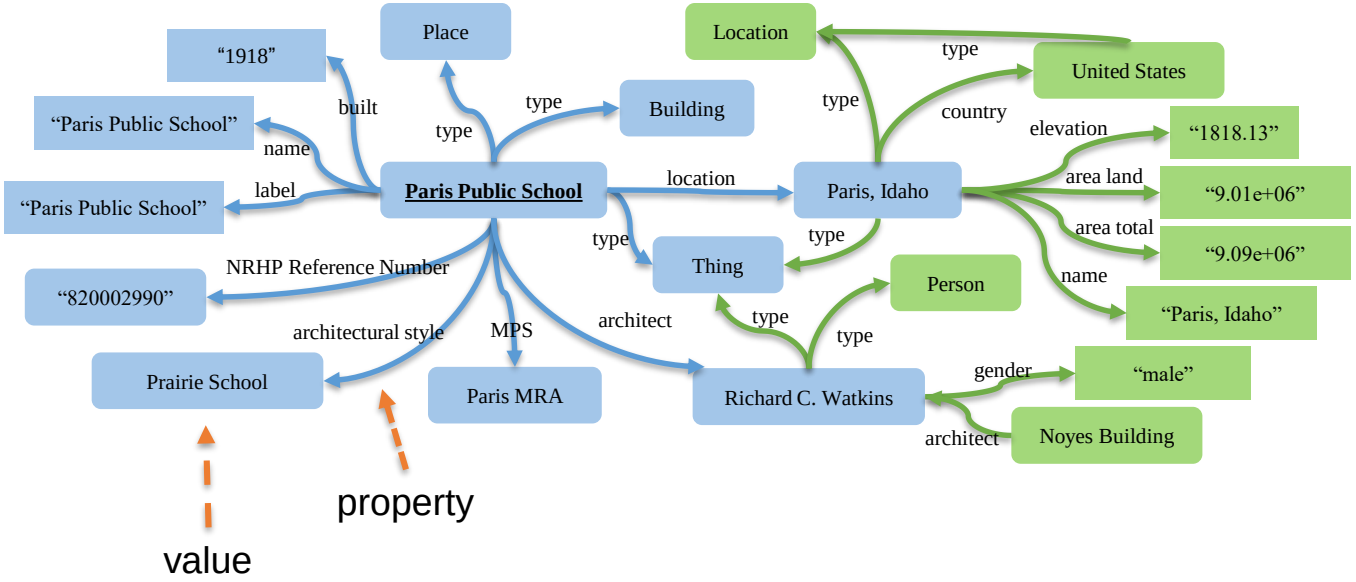
Outline

- Introduction
- ESSTER
  - Structural Importance
  - Textual Readability
  - Information Redundancy
  - Combinatorial Optimization
- Evaluation
- Conclusion

Introduction

RDF Graph

Entity Description



Entity Summary

Paris Public School
type: Building
location: Paris, Idaho
built: "1918"
architect: Richard C. Watkins
architectural style: Prairie School

Introduction

- Entity Description $Desc(e)$
  - Desc(e) = {<e, prop(t), val(t)>}
  - a set of triples in T, where e as subject or object

- Entity Summarization $Summ(e)$
  - Summ(e) ⊆ Desc(e) , |Summ(e)| ≤ k
  - Provide key information & compact
  - For human users ⇨ Reading Experience

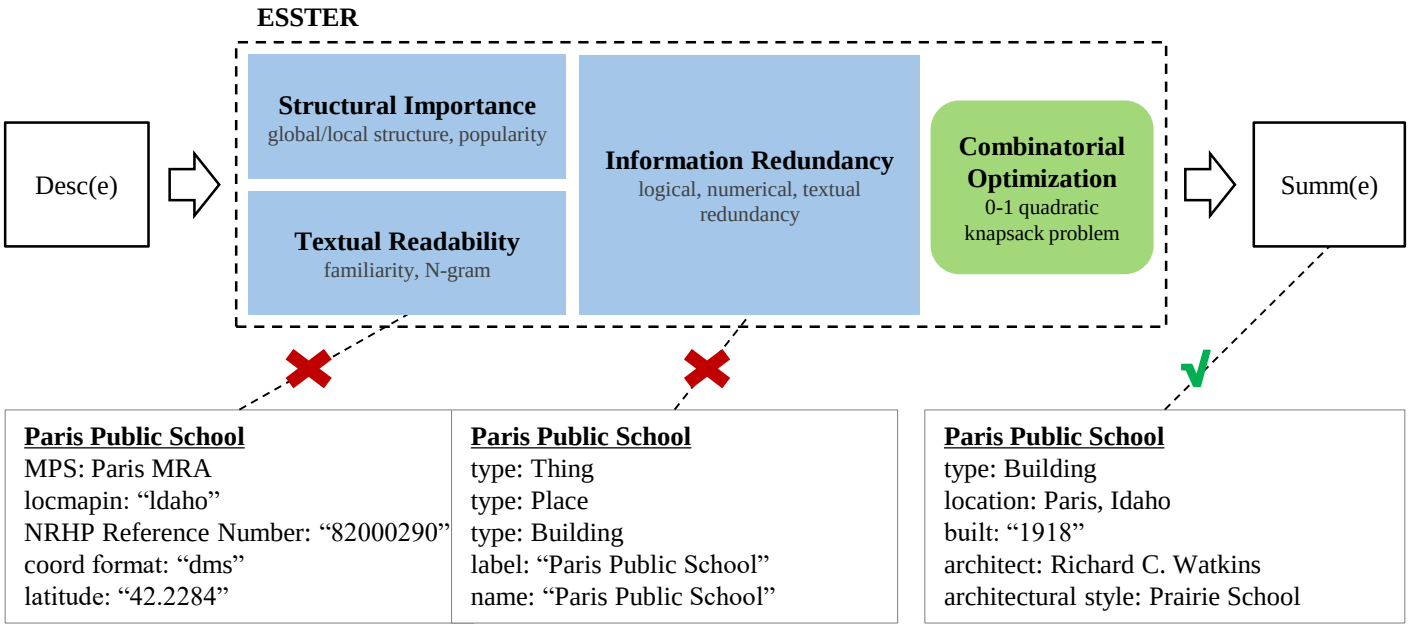
Paris Public School
MPS: Paris MRA
locmapin: “ldaho”
NRHP Reference Number: “82000290”
coord format: “dms”
latitude: “42.2284”

Paris Public School
type: Thing
type: Place
type: Building
label: “Paris Public School”
name: “Paris Public School”

Paris Public School
type: Building
location: Paris, Idaho
built: “1918”
architect: Richard C. Watkins
architectural style: Prairie School

ESSTER

Generating entity summaries of structural importance, high readability, and low redundancy.



Structural Importance

- Structural Importance
  - $W_{\text{struct}}(t) = \alpha \cdot \text{glb}(t) + (1 - \alpha) \cdot \text{loc}(t)$
- Global
  - $\text{glb}(t) = \text{ppop}_{\text{global}}(t) \cdot (1 - \text{vpop}(t))$
  - generality of property
  - characteristic of value
- Local
  - $\text{loc}(t) = (1 - \text{ppop}_{\text{local}}(t)) \cdot \text{vpop}(t)$
  - punishment on multi-valued properties
  - avoiding too technical/specific values

type: Thing
type: Architect ✓
hypernym: Architect

subject: American architects
subject: NRHP architects
birth year: “1858” ✓
wiki page ID: “34981613”

Textual Readability

location ✓

birth date ✓

country ✓

architectural style ✓

NRHP Reference Number ?

coord format ?

locmapin ?

MPS ?

Textual Readability

location ✓

birth date ✓

country ✓

architectural style ✓

NRHP Reference Number ?

coord format ?

locmapin ?

MPS ?

A property is familiar to users if
it is often used in an open-domain corpus

Textual Readability

location ✓ NRHP Reference Number ?

birth date ✓ coord format ?

country ✓ locmapin ?

architectural style ✓ MPS ?

B ： # documents in the corpus
$b(t)$ ： # documents where prop(t) appears
M ： # documents have been read by the user
m ： # documents have been read by the user and prop(t) appears

- Familiarity

$$Q(t) = \sum_{m=0}^{\min(b(t),M)} \frac{\binom{b(t)}{m} \cdot \binom{B-b(t)}{M-m}}{\binom{B}{M}} \cdot \texttt{familarity}(m),$$

$$\texttt{familarity}(m) = \frac{\log(m+1)}{\log(B+1)}.$$

- Textual Readability

$$W_{\text{text}}(t) = \log(Q(t)+1).$$

Information Redundancy

- Redundancy
  - $\mathrm{sim}(t_i, t_j)$

- Logical Redundancy
- Numerical Redundancy
- Textual Redundancy

Information Redundancy

- Redundancy
  - $\mathrm{sim}(t_i, t_j)$

- Logical Redundancy

  $\mathrm{sim}(t_i, t_j) = 1$ if
  - $\mathrm{prop}(t_i) = \mathrm{prop}(t_j) = \mathrm{rdf{:}\,type}$, $\mathrm{val}(t_i)$ and $\mathrm{val}(t_j)$ have rdfs: subClassOf relation
  - $\mathrm{val}(t_i) = \mathrm{val}(t_j)$, $\mathrm{prop}(t_i)$ and $\mathrm{prop}(t_j)$ have rdfs: subPropertyOf relation

type: Thing
type: Place
type: Building

label: “Paris Public School”
name: “Paris Public School”

Information Redundancy

- Redundancy
  - $\mathrm{sim}(t_i, t_j)$

- Numerical Redundancy

  $\mathrm{sim}(t_i, t_j) = \max\{\mathrm{sim}_p(t_i,\ t_j), \mathrm{sim}_v(t_i,\ t_j), 0\}$

  - $\mathrm{val}(t_i)$ and $\mathrm{val}(t_j)$ are both numerical values
  - $\mathrm{sim}_p(t_i, t_j) = \mathrm{ISub}(\mathrm{prop}(t_i), \mathrm{prop}(t_j))$
  - $\mathrm{sim}_v(t_i, t_j) = \begin{cases} -1, & \text{if } \mathrm{val}(t_i) \cdot \mathrm{val}(t_j) \le 0 \\ \frac{\min\{\mathrm{val}(t_i), \mathrm{val}(t_j)\}}{\max\{\mathrm{val}(t_i), \mathrm{val}(t_j)\}}, & \text{otherwise} \end{cases}$

area land: “9.01e+06” } **sim=0.9019**
area total: “9.09e+06” }

elevation: “1818.13” } **sim=0.0002**

Information Redundancy

- Redundancy
  - $\mathrm{sim}(t_i, t_j)$
- Textual Redundancy

  $\mathrm{sim}(t_i, t_j) = \max\{\mathrm{sim}_p(t_i,\ t_j), \mathrm{sim}_v(t_i,\ t_j), 0\}$

  - $\mathrm{sim}_p(t_i, t_j) = \mathrm{ISub}(\mathrm{prop}(t_i), \mathrm{prop}(t_j))$
  - $\mathrm{sim}_v(t_i, t_j) = \mathrm{ISub}(\mathrm{val}(t_i), \mathrm{val}(t_j))$

given name: “Richard”
name: “Richard C. Watkins” } **sim=0.78**

gender: “male” } **sim=0**

Combinatorial Optimization

- Goal: generate entity summary of
  - high structural importance $W_{\text{struct}}(t_i)$
  - high textual readability $W_{\text{text}}(t_i)$
  - low information redundancy $-\text{sim}(t_i, t_j)$
  - and satisfy size constraint *k*

Combinatorial Optimization

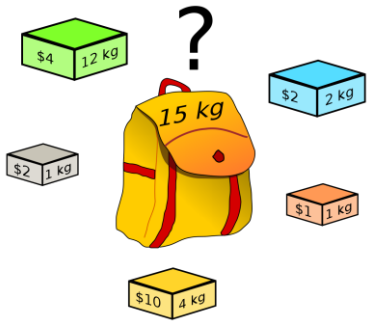
- 0-1 Quadratic Knapsack Problem (QKP)
  - max profit, satisfy weight constraint
  - profit:
    - structural importance, textual readability
    - low information redundancy
  - weight: x_i
    - 1 if $t_i \in \text{Summ(e)}$
    - 0 otherwise

$$\text{maximize} \sum_{i=1}^{|\text{desc}(e)|} \sum_{j=i}^{|\text{desc}(e)|} \text{profit}_{i,j} \cdot x_i \cdot x_j ,$$

$$\text{subject to} \sum_{i=1}^{|\text{desc}(e)|} x_i \le k ,$$

$$x_i \in \{0,1\} \text{ for all } i = 1 \ldots |\text{desc}(e)| .$$

$$\text{profit}_{i,j} = \begin{cases} (1-\delta) \cdot (W_{\text{struct}}(t_i) + W_{\text{text}}(t_i)) & i = j , \\ \delta \cdot (-\text{sim}(t_i, t_j)) & i \neq j , \end{cases}$$



Evaluation

- Data
  - ESBM v1.2
- Baselines
  - 9 unsupervised methods
- Results
  - top-2 on DBpedia under k=5
  - best in all the other three settings

F1 Scores

	DBpedia		LinkedMDB	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$
RELIN	0.242	0.455	0.203	0.258
DIVERSUM	0.249	0.507	0.207	0.358
FACES	0.270	0.428	0.169	0.263
FACES-E	0.280	0.488	0.313	0.393
CD	0.283	0.513	0.217	0.331
LinkSUM	0.287	0.486	0.140	0.279
BAFREC	**0.335**	0.503	0.360	0.402
KAFCA	0.314	0.509	0.244	0.397
MPSUM	0.314	0.512	0.272	0.423
ESSTER	0.324	**0.521**	**0.365**	**0.452**

Conclusion

- ESSTER: generating entity summaries by integrating
  - structural importance,
  - textual readability,
  - and information redundancy
  - via combinatorial optimization
- ESSTER achieves SOTA among unsupervised entity summarizers on ESBM v1.2
- Future Work
  - more powerful measures of readability and redundancy
  - incorporate these features into a neural network model

Thank you !

Questions ?