ESSTER at the EYRE 2020 Entity Summarization Task

Qingxia Liu[a], Gong Cheng[a] and Yuzhong Qu[a]

[a]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Abstract

Entity summaries provide human users with the key information about an entity. In this system paper, we present the implementation of our entity summarizer ESSTER. It aims at generating entity summaries that contain structurally important triples and exhibit high readability and low redundancy. For structural importance, we exploit the global and local characteristics of properties and values in RDF data. For readability, we learn the familarity of properties from a text corpus. To reduce redundancy, we perform logical reasoning and compute textual and numerical similarity between triples. ESSTER solves a combinatorial optimization problem to integrate these features. It achieves state-of-the-art results on the ESBM v1.2 dataset.

Keywords

Entity summarization, readability, redundancy

1. Introduction

In RDF data, an entity is described by a possibly large set (e.g., hundreds) of RDF triples. The entity summarization task is to automatically generate a compact summary to provide human users with the key information about an entity. Specifically, an entity summary is a size-constrained subset of triples selected from an entity description. Current methods [1, 2, 3, 4, 5, 6] are mainly focused on selecting important triples, but ignore the reading experience of human users. In this system paper, we present the implementation of our entity summarizer named **ESSTER** [7].[1] It aims at generating entity summaries of structural importance, high readability, and low redundancy. Improving textual readability and reducing information redundancy help to enhance the reading experience of users. Experiments on the ESBM v1.2 dataset [8] show that ESSTER achieves state-of-the-art results.

2. Task Definition

RDF data is a set of subject-predicate-object triples T. For an entity e, its description $\texttt{desc}(e)$ is the subset of triples in T such that e is the subject or the object. Each triple $t \in \texttt{desc}(e)$ provides a property-value pair $\langle p, v \rangle$ for e. When e is the subject of t, the property p is t's predicate and the value v is t's object. When e is the object of t, the property p is the inverse of t's predicate and the value v is t's subject. For convenience, we define $\texttt{prop}(t) = p$ and $\texttt{val}(t) = v$. Given an integer size constraint k, an entity summary S for e is a subset of $\texttt{desc}(e)$ satisfying $|S| \leq k$.

3. Implementation of ESSTER

ESSTER considers structural importance, readability, and redundancy. Below we present their computation and finally integrate them by solving a combinatorial optimization problem.

3.1. Structural Importance

We measure the structural importance of a triple t from two perspectives.

First, globally popular properties often reflect important aspects of entities, while globally unpopular values are informative. Therefore, we compute the global importance of a triple as follows:

$$\begin{aligned}\texttt{glb}(t) &= \texttt{ppop}_{\text{global}}(t) \cdot (1 - \texttt{vpop}(t)),\\ \texttt{ppop}_{\text{global}}(t) &= \frac{\log(\texttt{pfreq}_{\text{global}}(t) + 1)}{\log(|E| + 1)},\\ \texttt{vpop}(t) &= \frac{\log(\texttt{vfreq}(t) + 1)}{\log(|T| + 1)},\end{aligned} \tag{1}$$

where E is the set of all entities described in RDF data T, $\texttt{pfreq}_{\text{global}}(t)$ is the number of entity descriptions in T where $\texttt{prop}(t)$ appears, and $\texttt{vfreq}(t)$ is the number of triples in T where $\texttt{val}(t)$ is the value.

Second, multi-valued properties are intrinsically popular compared with single-valued properties. To compensate for this, we penalize multi-valued properties

Proceedings of the CIKM 2020 Workshops, October 19-20, 2020, Galway, Ireland

EMAIL: qxliu2013@smail.nju.edu.cn (Q. Liu); gcheng@nju.edu.cn (G. Cheng); yzqu@nju.edu.cn (Y. Qu)
URL: http://ws.nju.edu.cn/~gcheng (G. Cheng); http://ws.nju.edu.cn/~yzqu (Y. Qu)
ORCID: 0000-0001-6706-3776 (Q. Liu); 0000-0003-3539-7776 (G. Cheng); 0000-0003-2777-8149 (Y. Qu)

[1] https://github.com/nju-websoft/ESSTER

by using local popularity. We compute the local importance of a triple as follows:

$$\begin{aligned}\mathtt{loc}(t) &= (1 - \mathtt{ppop}_{\mathrm{local}}(t)) \cdot \mathtt{vpop}(t)\,,\\ \mathtt{ppop}_{\mathrm{local}}(t) &= \frac{\log(\mathtt{pfreq}_{\mathrm{local}}(t) + 1)}{\log(|\mathtt{desc}(e)| + 1)}\,,\end{aligned} \tag{2}$$

where $\mathtt{pfreq}_{\mathrm{local}}(t)$ is the number of triples in $\mathtt{desc}(e)$ where $\mathtt{prop}(t)$ is the property.

Finally, we compute structural importance:

$$W_{\mathrm{struct}}(t) = \alpha \cdot \mathtt{glb}(t) + (1 - \alpha) \cdot \mathtt{loc}(t)\,, \tag{3}$$

where $\alpha \in [0, 1]$ is a parameter to tune.

3.2. Textual Readability

To generate readable summaries, we measure the familiarity of a triple t based on its property $\mathtt{prop}(t)$. A property is familiar to users if it is often used in an open-domain corpus. Specifically, given a text corpus of B documents where M documents have been read by the user, let $b(t)$ be the number of documents where the name of $\mathtt{prop}(t)$ appears. We compute

$$\begin{aligned}Q(t) &= \sum_{m=0}^{\min(b(t),M)} \frac{\binom{b(t)}{m} \cdot \binom{B-b(t)}{M-m}}{\binom{B}{M}} \cdot \mathtt{familarity}(m)\,,\\ \mathtt{familarity}(m) &= \frac{\log(m + 1)}{\log(B + 1)}\,.\end{aligned} \tag{4}$$

Here, m represents the number of documents the user has read where the name of $\mathtt{prop}(t)$ appears, based on which $\mathtt{familarity}(m)$ gives the degree of familarity of $\mathtt{prop}(t)$ to the user. However, it is difficult to know m in practice, so $Q(t)$ computes the expected value of $\mathtt{familarity}(m)$. For simplicity, we assume M is a constant. In the experiments we set $M = 40$ and we use the Google Books Ngram[2] as our corpus.

Finally, we compute textual readability:

$$W_{\mathrm{text}}(t) = \log(Q(t) + 1). \tag{5}$$

3.3. Information Redundancy

To reduce redundancy in summaries, we measure the similarity between two triples t_i, t_j in various ways.

First, we perform logical reasoning to measure ontological similarity. We define $\mathtt{sim}(t_i, t_j) = 1$ if $\mathtt{prop}(t_i)$ and $\mathtt{prop}(t_j)$ are `rdf:type`, and `rdfs:subClassOf` is a relation between $\mathtt{val}(t_i)$ and $\mathtt{val}(t_j)$; or if $\mathtt{val}(t_i)$ and $\mathtt{val}(t_j)$ are equal, and `rdfs:subPropertyOf` is a relation between $\mathtt{prop}(t_i)$ and $\mathtt{prop}(t_j)$.

Otherwise, we rely on the similarity between properties and the similarity between values:

$$\mathtt{sim}(t_i, t_j) = \max\{\mathtt{sim}_{\mathrm{p}}(t_i, t_j), \mathtt{sim}_{\mathrm{v}}(t_i, t_j), 0\}\,, \tag{6}$$

where for $\mathtt{sim}_{\mathrm{p}}$ we use the ISub string similarity [9]. For $\mathtt{sim}_{\mathrm{v}}$, we differentiate between two cases.

In the first case, $\mathtt{val}(t_i)$ and $\mathtt{val}(t_j)$ are both numerical values. We compute

$$\mathtt{sim}_{\mathrm{v}}(t_i, t_j) = \begin{cases} -1 & \mathtt{val}(t_i) \cdot \mathtt{val}(t_j) \le 0\,,\\ \frac{\min\{\mathtt{val}(t_i),\mathtt{val}(t_j)\}}{\max\{\mathtt{val}(t_i),\mathtt{val}(t_j)\}} & \text{otherwise}\,. \end{cases} \tag{7}$$

In all other cases, we simply use ISub for $\mathtt{sim}_{\mathrm{v}}$.

3.4. Combinatorial Optimization

We formulate entity summarization as a 0-1 quadratic knapsack problem (QKP), and we solve it using a heuristic algorithm [10].

Specifically, we define the profit of choosing two triples t_i, t_j for a summary:

$$\mathtt{profit}_{i,j} = \begin{cases} (1 - \delta) \cdot (W_{\mathrm{struct}}(t_i) + W_{\mathrm{text}}(t_i)) & i = j\,,\\ \delta \cdot (-\mathtt{sim}(t_i, t_j)) & i \ne j\,, \end{cases} \tag{8}$$

where $\delta \in [0, 1]$ is a parameter to tune.

Finally, our goal is to

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{|\mathtt{desc}(e)|} \sum_{j=i}^{|\mathtt{desc}(e)|} \mathtt{profit}_{i,j} \cdot x_i \cdot x_j\,,\\ \text{subject to} \quad & \sum_{i=1}^{|\mathtt{desc}(e)|} x_i \le k\,,\\ & x_i \in \{0, 1\} \text{ for all } i = 1 \ldots |\mathtt{desc}(e)|\,. \end{aligned} \tag{9}$$

4. Experiments

4.1. Settings

We use the ESBM v1.2 dataset [8]. It provides ground-truth summaries under $k = 5$ and $k = 10$ for entities in DBpedia and LinkedMDB. We follow the provided training-development-test splits for 5-fold cross validation, and we use the training and development sets for tuning our parameters α and δ by grid search in the range of 0–1 with 0.01 increments. We use F1 score as the evaluation metric.

[2] http://books.google.com/ngrams

Table 1
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**

4.2. Results

Table 1 presents the evaluation results. We compare with known results of existing unsupervised entity summarizers [8]. On DBpedia under $k = 5$, BAFREC [6] achieves the highest F1 score, and is closely followed by ESSTER. In all the other three settings, ESSTER outperforms all the baselines. Overall, ESSTER achieves state-of-the-art results on ESBM v1.2.

5. Conclusion

In this system paper, we presented the implementation of our entity summarizer ESSTER. By integrating structural importance, textual readability, and information redundancy via combinatorial optimization, ESSTER achieves state-of-the-art results among unsupervised entity summarizers on the ESBM v1.2 dataset. However, the results are not comparable with supervised neural entity summarizers [11, 12].

For the future work, we will consider more powerful measures of readability and redundancy, and will incorporate these features into a neural network model.

Acknowledgments

This work was supported by the National Key R&D Program of China (2018YFB1004300) and by the NSFC (61772264).

References

[1] Q. Liu, G. Cheng, K. Gunaratna, Y. Qu, Entity summarization: State of the art and future challenges, CoRR abs/1910.08252 (2019).

[2] G. Cheng, T. Tran, Y. Qu, RELIN: relatedness and informativeness-based centrality for entity summarization, in: ISWC'11, Part I, 2011, pp. 114–129. doi:10.1007/978-3-642-25073-6_8.

[3] K. Gunaratna, K. Thirunarayan, A. P. Sheth, FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering, in: AAAI'15, 2015, pp. 116–122.

[4] K. Gunaratna, K. Thirunarayan, A. P. Sheth, G. Cheng, Gleaning types for literals in RDF triples with application to entity summarization, in: ESWC'16, 2016, pp. 85–100. doi:10.1007/978-3-319-34129-3_6.

[5] A. Thalhammer, N. Lasierra, A. Rettinger, LinkSUM: Using link analysis to summarize entity data, in: ICWE'16, 2016, pp. 244–261. doi:10.1007/978-3-319-38791-8_14.

[6] H. Kroll, D. Nagel, W.-T. Balke, BAFREC: Balancing frequency and rarity for entity characterization in linked open data, in: EYRE'18, 2018.

[7] Q. Liu, G. Cheng, Y. Qu, Entity summarization with high readability and low redundancy, Sci. Sin. Inform. 50 (2020) 845–861. doi:10.1360/SSI-2019-0291.

[8] Q. Liu, G. Cheng, K. Gunaratna, Y. Qu, ESBM: an entity summarization benchmark, in: ESWC'20, 2020, pp. 548–564. doi:10.1007/978-3-030-49461-2_32.

[9] G. Stoilos, G. B. Stamou, S. D. Kollias, A string metric for ontology alignment, in: ISWC'05, 2005, pp. 624–637. doi:10.1007/11574620_45.

[10] Z. Yang, G. Wang, F. Chu, An effective GRASP and tabu search for the 0-1 quadratic knapsack problem, Comput. Oper. Res. 40 (2013) 1176–1185. doi:10.1016/j.cor.2012.11.023.

[11] Q. Liu, G. Cheng, Y. Qu, Deeplens: Deep learning for entity summarization, in: DL4KG'20, 2020.

[12] J. Li, G. Cheng, Q. Liu, W. Zhang, E. Kharlamov, K. Gunaratna, H. Chen, Neural entity summarization with joint encoding and weak supervision, in: IJCAI'20, 2020, pp. 1644–1650. doi:10.24963/ijcai.2020/228.