ESBM: An Entity Summarization BenchMark

Qingxia Liu[1], Gong Cheng[1(✉)], Kalpa Gunaratna[2], and Yuzhong Qu[1]

[1] National Key Laboratory for Novel Software Technology, Nanjing University,
Nanjing, China
qxliu2013@smail.nju.edu.cn, {gcheng,yzqu}@nju.edu.cn
[2] Samsung Research America, Mountain View, CA, USA
k.gunaratna@samsung.com

Abstract. Entity summarization is the problem of computing an optimal compact summary for an entity by selecting a size-constrained subset of triples from RDF data. Entity summarization supports a multiplicity of applications and has led to fruitful research. However, there is a lack of evaluation efforts that cover the broad spectrum of existing systems. One reason is a lack of benchmarks for evaluation. Some benchmarks are no longer available, while others are small and have limitations. In this paper, we create an Entity Summarization BenchMark (ESBM) which overcomes the limitations of existing benchmarks and meets standard desiderata for a benchmark. Using this largest available benchmark for evaluating general-purpose entity summarizers, we perform the most extensive experiment to date where 9 existing systems are compared. Considering that all of these systems are unsupervised, we also implement and evaluate a supervised learning based system for reference.

Keywords: Entity summarization · Triple ranking · Benchmarking

1 Introduction

RDF data describes entities with triples representing property values. In an RDF dataset, the description of an entity comprises all the RDF triples where the entity appears as the subject or the object. An example entity description is shown in Fig. 1. Entity descriptions can be large. An entity may be described in dozens or hundreds of triples, exceeding the capacity of a typical user interface. A user served with all of those triples may suffer information overload and find it difficult to quickly identify the small set of triples that are truly needed. To solve the problem, an established research topic is *entity summarization* [15], which aims to compute an optimal compact summary for the entity by selecting a size-constrained subset of triples. An example entity summary under the size constraint of 5 triples is shown in the bottom right corner of Fig. 1.

Description of Tim Berners-Lee:

<Tim Berners Lee, alias, "TimBL">
<Tim Berners Lee, name, "Tim Berners-Lee">
<Tim Berners Lee, givenName, "Tim">
<Tim Berners Lee, birthYear, "1955">
<Tim Berners Lee, birthDate, "1955-06-08">
<Tim Berners Lee, birthPlace, England>
<Tim Berners Lee, birthPlace, London>
<Tim Berners Lee, type, People Educated At Emanuel School>
<Tim Berners Lee, type, Scientist>
<Tim Berners Lee, type, Living People>
<Tim Berners Lee, type, Person>
<Tim Berners-Lee, type, Agent>
<Tim Berners-Lee, award, Royal Society>
<Tim Berners-Lee, award, Royal Academy of Engineering >
<Tim Berners-Lee, award, Order of Merit>
<Tim Berners-Lee, award, Royal Order of the British Empire>
<Tim Berners-Lee, spouse, Rosemary Leith>

<Tim Berners-Lee, child, Ben Berners-Lee>
<Tim Berners-Lee, child, Alice Berners-Lee>
<Conway Berners-Lee, child, Tim Berners-Lee>
<Weaving the Web, author, Tim Berners-Lee>
<Tabulator, author, Tim Berners-Lee>
<Paul Otlet, influenced, Tim Berners-Lee>
<John Postel, influenced, Tim Berners-Lee>
<World Wide Web, developer, Tim Berners-Lee>
<World Wide Web Foundation, foundedBy, Tim Berners-Lee>
<World Wide Web Foundation, keyPerson, Tim Berners-Lee>

Summary:

<Tim Berners Lee, birthDate, "1955-06-08">
<Tim Berners Lee, birthPlace, England>
<Tim Berners-Lee, type, Scientist>
<Tim Berners-Lee, award, Royal Society>
<World Wide Web, developer, Tim Berners-Lee>

Fig. 1. Description of entity `Tim Berners-Lee` and a summary thereof.

Entity summarization supports a multiplicity of applications [6,21]. Entity summaries constitute entity cards displayed in search engines [9], provide background knowledge for enriching documents [26], and facilitate research activities with humans in the loop [3,4]. This far-reaching application has led to fruitful research as reviewed in our recent survey paper [15]. Many entity summarizers have been developed, most of which generate summaries for general purposes.

Research Challenges. However, two challenges face the research community. First, there is a *lack of benchmarks* for evaluating entity summarizers. As shown in Table 1, some benchmarks are no longer available. Others are available [7,8,22] but they are small and have limitations. Specifically, [22] has a task-specific nature, and [7,8] exclude classes and/or literals. These benchmarks could not support a comprehensive evaluation of general-purpose entity summarizers. Second, there is a *lack of evaluation efforts* that cover the broad spectrum of existing systems to compare their performance and assist practitioners in choosing solutions appropriate to their applications.

Contributions. We address the challenges with two contributions. First, we create an **E**ntity **S**ummarization **B**ench**M**ark (ESBM) which overcomes the limitations of existing benchmarks and meets the desiderata for a successful benchmark [18]. ESBM has been published on GitHub with extended documentation and a permanent identifier on `w3id.org`[1] under the ODC-By license. As the largest available benchmark for evaluating general-purpose entity summarizers, ESBM contains 175 heterogeneous entities sampled from two datasets, for which 30 human experts create 2,100 general-purpose ground-truth summaries under two size constraints. Second, using ESBM, we evaluate 9 existing general-purpose entity summarizers. It represents the most extensive evaluation effort to date. Considering that existing systems are unsupervised, we also implement and evaluate a supervised learning based entity summarizer for reference.

In this paper, for the first time we comprehensively describe the creation and use of ESBM. We report ESBM v1.2—the latest version, while early

---

[1] https://w3id.org/esbm.

Table 1. Existing benchmarks for evaluating entity summarization.

	Dataset	Number of entities	Availability
WhoKnows?Movies! [22]	Freebase	60	**Available**[a]
Langer et al. [13]	DBpedia	14	Unavailable
FRanCo [1]	DBpedia	265	Unavailable
Benchmark for evaluating RELIN [2]	DBpedia	149	Unavailable
Benchmark for evaluating DIVERSUM [20]	IMDb	20	Unavailable
Benchmark for evaluating FACES [7]	DBpedia	50	**Available**[b]
Benchmark for evaluating FACES-E [8]	DBpedia	80	**Available**[b]

[a] http://yovisto.com/labs/iswc2012
[b] http://wiki.knoesis.org/index.php/FACES

versions have successfully supported the entity summarization shared task at the EYRE 2018 workshop[2] and the EYRE 2019 workshop.[3] We will also educate on the use of ESBM at an ESWC 2020 tutorial on entity summarization[4].

The remainder of the paper is organized as follows. Section 2 reviews related work and limitations of existing benchmarks. Section 3 describes the creation of ESBM, which is analyzed in Sect. 4. Section 5 presents our evaluation. In Sect. 6 we discuss limitations of our study and perspectives for future work.

2   Related Work

We review methods and evaluation efforts for entity summarization.

Methods for Entity Summarization. In a recent survey [15] we have categorized the broad spectrum of research on entity summarization. Below we briefly review *general-purpose* entity summarizers which mainly rely on generic technical features that can apply to a wide range of domains and applications. We will not address methods that are domain-specific (e.g., for movies [25] or timelines [5]), task-specific (e.g., for facilitating entity resolution [3] or entity linking [4]), or context-aware (e.g., contextualized by a document [26] or a query [9]).

RELIN [2] uses a weighted PageRank model to rank triples according to their statistical informativeness and relatedness. DIVERSUM [20] ranks triples by property frequency and generates a summary with a strong constraint that avoids selecting triples having the same property. SUMMARUM [24] and LinkSUM [23] mainly rank triples by the PageRank scores of property values that are entities. LinkSUM also considers backlinks from values. FACES [7], and its extension FACES-E [8] which adds support for literals, cluster triples by their bag-of-words based similarity and choose top-ranked triples from as many different clusters as

possible. Triples are ranked by statistical informativeness and property value frequency. CD [28] models entity summarization as a quadratic knapsack problem that maximizes the statistical informativeness of the selected triples and in the meantime minimizes the string, numerical, and logical similarity between them. In ES-LDA [17], ES-LDA$_{ext}$ [16], and MPSUM [27], a Latent Dirichlet Allocation (LDA) model is learned where properties are treated as topics, and each property is a distribution over all the property values. Triples are ranked by the probabilities of properties and values. MPSUM further avoids selecting triples having the same property. BAFREC [12] categorizes triples into meta-level and data-level. It ranks meta-level triples by their depths in an ontology and ranks data-level triples by property and value frequency. Triples having textually similar properties are penalized to improve diversity. KAFCA [11] ranks triples by the depths of properties and values in a hierarchy constructed by performing the Formal Concept Analysis (FCA). It tends to select triples containing infrequent properties but frequent values, where frequency is computed at the word level.

Limitations of Existing Benchmarks. For evaluating entity summarization, compared with task completion based *extrinsic evaluation*, ground truth based *intrinsic evaluation* is more popular because it is easy to perform and the results are reproducible. Its idea is to create a benchmark consisting of human-made ground-truth summaries, and then compute how much a machine-generated summary is close to a ground-truth summary.

Table 1 lists known benchmarks, including dedicated benchmarks [1,13,22] and those created for evaluating a particular entity summarizer [2,7,8,20]. It is not surprising that these benchmarks are not very large since it is expensive to manually create high-quality summaries for a large set of entities. Unfortunately, some of these benchmarks are not publicly available at this moment. Three are available [7,8,22] but they are relatively small and have limitations. Specifically, WhoKnows?Movies! [22] is not a set of ground-truth summaries but annotates each triple with the ratio of movie questions that were correctly answered based on that triple, as an indicator of its importance. This kind of task-specific ground truth may not be suitable for evaluating general-purpose entity summarizers. The other two available benchmarks were created for evaluating FACES/-E [7,8]. Classes and/or literals are not included because they could not be processed by FACES/-E and hence were filtered out. Such benchmarks could not comprehensively evaluate most of the existing entity summarizers [2,11,12,20,27,28] that can handle classes and literals. These limitations of available benchmarks motivated us to create a new ground truth consisting of *general-purpose summaries* for a *larger set of entities* involving *more comprehensive triples* where property values can be entities, classes, or literals.

3   Creating ESBM

To overcome the above-mentioned limitations of existing benchmarks, we created a new benchmark called ESBM. To date, it is the largest available benchmark for evaluating general-purpose entity summarizers. In this section, we will first

specify our design goals. Then we describe the selection of entity descriptions and the creation of ground-truth summaries. We partition the data to support cross-validation for parameter fitting. Finally we summarize how our design goals are achieved and how ESBM meets standard desiderata for a benchmark.

3.1 Design Goals

The creation of ESBM has two main design goals. First, a successful benchmark should meet seven desiderata [18]: accessibility, affordability, clarity, relevance, solvability, portability, and scalability, which we will detail in Sect. 3.5. Our design of ESBM aims to satisfy these basic requirements. Second, in Sect. 2 we discussed the limitations of available benchmarks, including task specificness, small size, and triple incomprehensiveness. Besides, all the existing benchmarks use a single dataset and hence may weaken the generalizability of evaluation results. We aim to overcome these limitations when creating ESBM. In Sect. 3.5 we will summarize how our design goals are achieved.
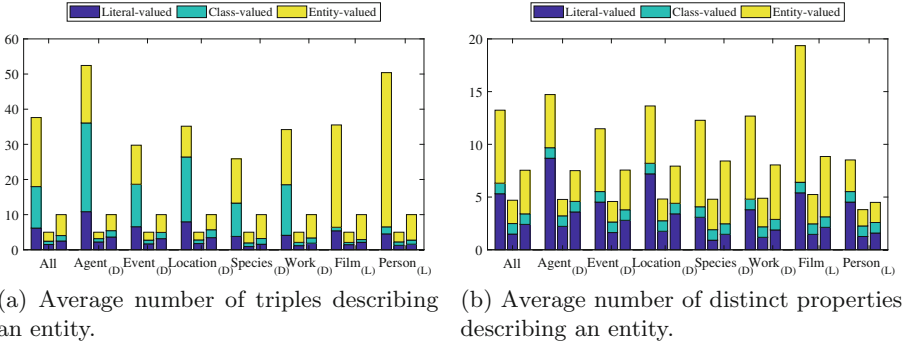
3.2 Entity Descriptions

To choose entity descriptions to summarize, we sample entities from selected datasets and filter their triples. The process is detailed below.

Datasets. We sample entities from two datasets of different kinds: an encyclopedic dataset and a domain-specific dataset. For the encyclopedic dataset we choose DBpedia [14], which has been used in other benchmarks [1,2,7,8,13]. We use the English version of DBpedia 2015-10[5]—the latest version when we started to create ESBM. For the domain-specific dataset we choose LinkedMDB [10], which is a popular movie database. The movie domain is also the focus of some existing benchmarks [20,22] possibly because this domain is familiar to the lay audience so that it would be easy to find qualified human experts to create ground-truth summaries. We use the latest available version of LinkedMDB.[6]

Entities. For DBpedia we sample entities from five large classes: `Agent`, `Event`, `Location`, `Species`, and `Work`. They collectively contain 3,501,366 entities (60%) in the dataset. For LinkedMDB we sample from `Film` and `Person`, which contain 159,957 entities (24%) in the dataset. Entities from different classes are described by very different properties as we will see in Sect. 4.3, and hence help to assess the generalizability of an entity summarizer. According to the human efforts we could afford, from each class we randomly sample 25 entities. The total number of selected entities is 175. Each selected entity should be described in at least 20 triples so that summarization would not be a trivial task. This requirement follows common practice in the literature [1,2,7,20] where a minimum constraint in the range of 10–20 was posed.

(a) Average number of triples describing an entity.

(b) Average number of distinct properties describing an entity.

Fig. 2. Composition of entity descriptions (the left bar in each group), top-5 ground-truth summaries (the middle bar), and top-10 ground-truth summaries (the right bar), grouped by class in DBpedia (D) and LinkedMDB (L).

Triples. For DBpedia, entity descriptions comprise triples in the following dump files: *instance types*, *instance types transitive*, *YAGO types*, *mappingbased literals*, *mappingbased objects*, *labels*, *images*, *homepages*, *persondata*, *geo coordinates mappingbased*, and *article categories*. We do not import dump files that provide metadata about Wikipedia articles such as *page links* and *page length*. We do not import *short abstracts* and *long abstracts* as they provide handcrafted textual entity summaries; it would be inappropriate to include them in a benchmark for evaluating entity summarization. For LinkedMDB we import all the triples in the dump file except `sameAs` links which do not express facts about entities but are of more technical nature. Finally, as shown in Fig. 2a (the left bar in each group), the mean number of triples in an entity description is in the range of 25.88–52.44 depending on the class, and the overall mean value is 37.62.

3.3 Ground-Truth Summaries

We invite 30 researchers and students to create ground-truth summaries for entity descriptions. All the participants are familiar with RDF.

Task Assignment. Each participant is assigned 35 entities consisting of 5 entities randomly selected from each of the 7 classes in ESBM. The assignment is controlled to ensure that each entity in ESBM is processed by 6 participants. A participant creates two summaries for each entity description by selecting different numbers of triples: a *top-5 summary* containing 5 triples, and a *top-10 summary* containing 10 triples. Therefore, we will be able to evaluate entity summarizers under different size constraints. The choice of these two numbers follows previous work [2,7,8]. Participants work independently and they may create different summaries for an entity. It is not feasible to ask participants to reach an agreement. It is also not reasonable to merge different summaries into a single version. So we keep different summaries and will use all of them in the evaluation. The total number of ground-truth summaries is $175 \cdot 6 \cdot 2 = 2,100$.

Fig. 3. User interface for creating ground-truth entity summaries.

Procedure. Participants are instructed to create *general-purpose summaries* that are not specifically created for any particular task. They read and select triples using a Web-based user interface shown in Fig. 3. All the triples in an entity description are listed in random order but those having a common property are placed together for convenient reading and comparison. For IRIs, their human-readable labels (`rdfs:label`) are shown if available. To help participants understand a property value that is an unfamiliar entity, a click on it will open a pop-up showing a short textual description extracted from the first paragraph of its Wikipedia/IMDb page. Any triple can be selected into the top-5 summary, the top-10 summary, or both. The top-5 summary is not required to be a subset of the top-10 summary.

3.4 Training, Validation, and Test Sets

Some entity summarizers need to tune hyperparameters or fit models. To make their evaluation results comparable with each other, we specify a split of our data into training, validation, and test sets. We provide a partition of the 175 entities in ESBM into 5 equally sized subsets $P_0, \ldots, P_4$ to support 5-fold cross-validation. Entities of each class are partitioned evenly among the subsets. For $0 \leq i \leq 4$, the i-th fold uses $P_i, P_{i+1 \bmod 5}, P_{i+2 \bmod 5}$ as the training set (e.g., for model fitting), uses $P_{i+3 \bmod 5}$ for validation (e.g., tuning hyperparameters), and retains $P_{i+4 \bmod 5}$ as the test set. Evaluation results are averaged over the 5 folds.

3.5 Conclusion

ESBM overcomes the limitations of available benchmarks discussed in Sect. 2. It contains 175 entities which is 2–3 times as large as available benchmarks [7,8,22].

In ESBM, property values are not filtered as in [7,8] but can be any entity, class, or literal. Different from the task-specific nature of [22], ESBM provides general-purpose ground-truth summaries for evaluating general-purpose entity summarizers.

Besides, ESBM meets the seven desiderata proposed in [18] as follows.

– **Accessibility.** ESBM is publicly available and has a permanent identifier on w3id.org.
– **Affordability.** ESBM is with an open-source program and example code for evaluation. The cost of using ESBM is minimized.
– **Clarity.** ESBM is documented clearly and concisely.
– **Relevance.** ESBM samples entities from two real datasets that have been widely used. The summarization tasks are natural and representative.
– **Solvability.** An entity description in ESBM has at least 20 triples and a mean number of 37.62 triples, from which 5 or 10 triples are to be selected. The summarization tasks are not trivial and not too difficult.
– **Portability.** ESBM can be used to evaluate any general-purpose entity summarizer that can process RDF data.
– **Scalability.** ESBM samples 175 entities from 7 classes. It is reasonably large and diverse to evaluate mature entity summarizers but is not too large to evaluate research prototypes.

However, ESBM has its own limitations, which we will discuss in Sect. 6.

4   Analyzing ESBM

In this section, we will first characterize ESBM by providing some basic statistics and analyzing the triple composition and heterogeneity of entity descriptions. Then we compute inter-rater agreement to show how much consensus exists in the ground-truth summaries given by different participants.

4.1   Basic Statistics

The 175 entity descriptions in ESBM collectively contain 6,584 triples, of which 37.44% are selected into at least one top-5 summary and 58.15% appear in at least one top-10 summary, showing a wide selection by the participants. However, many of them are selected only by a single participant; 20.46% and 40.23% are selected by different participants into top-5 and top-10 summaries, respectively. We will further analyze inter-rater agreement in Sect. 4.4.

We calculate the overlap between the top-5 and the top-10 summaries created by the same participant for the same entity. The mean overlap is in the range of 4.80–4.99 triples depending on the class, and the overall mean value is 4.91, showing that the top-5 summary is usually a subset of the top-10 summary.

4.2   Triple Composition

In Fig. 2 we present the composition of entity descriptions (the left bar in each group) and their ground-truth summaries (the middle bar for top-5 and the right bar for top-10) in ESBM, in terms of the average number of triples describing an entity (Fig. 2a) and in terms of the average number of distinct properties describing an entity (Fig. 2b). Properties are divided into literal-valued, class-valued, and entity-valued. Triples are divided accordingly.

In Fig. 2a, both class-valued and entity-valued triples occupy a considerable proportion of the entity descriptions in DBpedia. Entity-valued triples predominate in LinkedMDB. Literal-valued triples account for a small proportion in both datasets. However, they constitute 30% in top-5 ground-truth summaries and 25% in top-10 summaries. Entity summarizers that cannot process literals [7,17,23,24] have to ignore these notable proportions, thereby significantly influencing their performance.
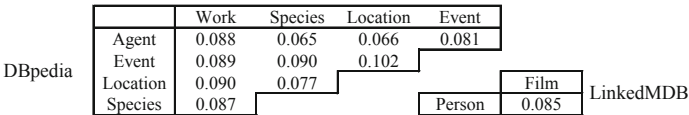
	Work	Species	Location	Event			
Agent	0.088	0.065	0.066	0.081			
Event	0.089	0.090	0.102				
Location	0.090	0.077				Film	LinkedMDB
Species	0.087				Person	0.085	

DBpedia

Fig. 4. Jaccard similarity between property sets describing different classes.

Table 2. Popular properties in ground-truth summaries.

In top-5 summaries							In top-10 summaries						
Agent	Event	Location	Species	Work	Film	Person	Agent	Event	Location	Species	Work	Film	Person
type	type	type	type	type	director	type	type	type	type	family	type	director	type
birthDate	date	country	family		type	actor	subject	subject	country	type	subject	actor	actor
							birthDate	date	subject	order	genre	type	label
							label			class		writer	page
										genus		producer	
										subject		date	
										kingdom		language	

In Fig. 2b, in terms of distinct properties, entity-valued and literal-valued triples have comparable numbers in entity descriptions since many entity-valued properties are multi-valued. Specifically, an entity is described by 13.24 distinct properties, including 5.31 literal-valued (40%) and 6.93 entity-valued (52%). Multi-valued properties appear in every entity description and they constitute 35% of the triples. However, in top-5 ground-truth summaries, the average number of distinct properties is 4.70 and is very close to 5, indicating that the participants are not inclined to select multiple values of a property. Entity summarizers that prefer diverse properties [7,8,12,20,27,28] may exhibit good performance.

4.3   Entity Heterogeneity

Entities from different classes are described by different sets of properties. For each class we identify the set of properties describing at least one entity from the class. The Jaccard similarity between properties sets for each pair of classes is very low, as shown in Fig. 4. Such heterogeneous entity descriptions help to assess the generalizability of an entity summarizer.

Table 2 shows popular properties that appear in at least 50% of the ground-truth summaries for each class. Some universal properties like `rdf:type` and `dct:subject` are popular for most classes. We also see class-specific properties, e.g., `dbo:birthDate` for `Agent`, `dbo:family` for `Species`. However, the results suggest that it would be unrealistic to generate good summaries by manually selecting properties for each class. For example, among 13.24 distinct properties describing an entity, only 1–2 are popular in top-5 ground-truth summaries. The importance of properties is generally contextualized by concrete entities.

4.4   Inter-rater Agreement

Recall that each entity in ESBM has six top-5 ground-truth summaries and six top-10 summaries created by different participants. We calculate the average overlap between these summaries in terms of the number of common triples they contain. As shown in Table 3, the results are generally comparable with those reported for other benchmarks in the literature. There is a moderate degree of agreement between the participants.

Table 3. Inter-rater agreement.

	ESBM	[2]	[7]	[8]
Overlap between top-5 summaries	1.99 (39.8%)	2.91 (58.2%)	1.92 (38.4%)	2.12 (42.4%)
Overlap between top-10 summaries	5.42 (54.2%)	7.86 (78.6%)	4.64 (46.4%)	5.44 (54.4%)
Ground-truth summaries per entity	6	4.43	≥ 7	≥ 4

5   Evaluating with ESBM

We used ESBM to perform the most extensive evaluation of general-purpose entity summarizers to date. In this section, we will first describe evaluation criteria. Then we introduce the entity summarizers that we evaluate. Finally we present evaluation results.

5.1   Evaluation Criteria

Let S_m be a machine-generated entity summary. Let S_h be a human-made ground-truth summary. To compare S_m with S_h and assess the quality of S_m based on how much S_m is close to S_h, it is natural to compute precision (P), recall (R), and F1. The results are in the range of 0–1:

$$\mathrm{P} = \frac{|S_m \cap S_h|}{|S_m|}, \quad \mathrm{R} = \frac{|S_m \cap S_h|}{|S_h|}, \quad \mathrm{F1} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \tag{1}$$

In the experiments we configure entity summarizers to output at most k triples and we set $k = |S_h|$, i.e., $k = 5$ and $k = 10$ are our two settings corresponding to the sizes of ground-truth summaries. We will trivially have P=R=F1 if $|S_m| = |S_h|$. However, some entity summarizers may output less than k triples. For example, DIVERSUM [20] disallows an entity summary to contain triples having the same property. It is possible that an entity description contains less than k distinct properties and hence DIVERSUM has to output less than k triples. In this case, $\mathrm{P} \neq \mathrm{R}$ and one should rely on F1.

In the evaluation, for each entity in ESBM, we compare a machine-generated summary with each of the 6 ground-truth summaries by calculating F1, and take their aggregation value. Finally we report the mean F1 over all the entities. For aggregation function, we report the results of average, to show an overall match with all the different ground truths; on the website we also give the results of maximum, to show the best match with each individual ground truth.

5.2   Participating Entity Summarizers

We not only evaluate existing entity summarizers but also compare them with two special entity summarizers we create: an oracle entity summarizer which is used to show the best possible performance on ESBM, and a new supervised learning based entity summarizer.

Existing Entity Summarizers. We evaluate 9 out of the 12 general-purpose entity summarizers reviewed in Sect. 2. We re-implement RELIN [2], DIVER-SUM [20], LinkSUM [23], FACES [7], FACES-E [8], and CD [28], while MPSUM [27], BAFREC [12], and KAFCA [11] are open source. We exclude SUMMARUM [24], ES-LDA [17], and ES-LDA$_{ext}$ [16] because LinkSUM represents an extension of SUMMARUM, and MPSUM represents an extension of ES-LDA and ES-LDA$_{ext}$.

We follow the original implementation and suggested configuration of existing entity summarizers as far as possible. However, for RELIN, we replace its Google-based relatedness measure with a string metric [19] because Google's search API is no longer free. We also use this metric to replace the unavailable UMBC's SimService used in FACES-E. For DIVERSUM, we ignore its witness count measure since it does not apply to ESBM. For LinkSUM, we obtain backlinks between entities in LinkedMDB via their corresponding entities in DBpedia.

RELIN, CD, and LinkSUM compute a weighted combination of two scoring components. We tune these hyperparameters in the range of 0–1 in 0.01 increments. Since these summarizers are unsupervised, we use both the training set and the validation set described in Sect. 3.4 for tuning hyperparameters.

Oracle Entity Summarizer. We implement an entity summarizer denoted by ORACLE to approximate the best possible performance on ESBM and form a reference point used for comparisons. ORACLE simply outputs k triples that are selected by the most participants into ground-truth summaries.

Supervised Learning Based Entity Summarizer. Existing general-purpose entity summarizers are unsupervised. We implement a supervised learning based entity summarizer with features that are used by existing entity summarizers. A triple with property p and value v describing entity e is represented by the following features:

- $\mathtt{gf}_\mathbb{T}$: the number of triples in the dataset where p appears [12,23],
- $\mathtt{lf}$: the number of triples in the description of e where p appears [20,23],
- $\mathtt{vf}_\mathbb{T}$: the number of triples in the dataset where v appears [7,8,12], and
- $\mathtt{si}$: the self-information of the triple [2,7,8,28].

We also add three binary features:

- $\mathtt{isC}$: whether v is a class,
- $\mathtt{isE}$: whether v is an entity, and
- $\mathtt{isL}$: whether v is a literal.

Based on the training and validation sets described in Sect. 3.4, we implement and tune 6 pointwise learning to rank models provided by Weka: SMOreg, LinearRegression, MultilayerPerceptron, AdditiveRegression, REPTree, and RandomForest. Each model outputs k top-ranked triples as a summary.

5.3   Evaluation Results

We first report the overall evaluation results to show which entity summarizer generally performs better. Then we break down the results into different entity types (i.e., classes) for detailed comparison. Finally we present and analyze the performance of our supervised learning based entity summarizer.

Overall Results of Existing Entity Summarizers. Table 4 presents the results of all the participating entity summarizers on two datasets under two size constraints. We compare nine existing summarizers using one-way ANOVA post-hoc LSD and we show whether the difference between each pair of them is statistical significant at the 0.05 level. Among existing summarizers, BAFREC achieves the highest F1 under $k = 5$. It significantly outperforms six existing summarizers on DBpedia and outperforms all the eight ones on LinkedMDB. It is also among the best under $k = 10$. MPSUM follows BAFREC under $k = 5$ but performs slightly better under $k = 10$. Other top-tier results belong to KAFCA on DBpedia and FACES-E on LinkedMDB.

Table 4. Average F1 over all the entities in a dataset. For the nine existing entity summarizers, significant improvements and losses over each other are indicated by ▲ and ▼ ($p < 0.05$), respectively. Insignificant differences are indicated by ○.

	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 ▲▲▲○▲▲○○-
ORACLE	0.595	0.713	0.619	0.678
SMOreg	0.279	0.543	0.403	0.472
LinearRegression	0.319	0.556	0.401	0.471
MultilayerPerceptron	0.340	0.560	0.390	0.477
AdditiveRegression	0.345	0.558	0.415	0.510
REPTree	0.392	0.570	0.455	0.538
RandomForest	0.399	0.576	0.449	0.506

The F1 scores of ORACLE are in the range of 0.595–0.713. It is impossible for ORACLE or any other summarizer to reach F1 = 1, because for each entity in ESBM there are six ground-truth summaries which are often different and hence cannot simultaneously match a machine-generated summary. However, the gap between the results of ORACLE and the best results of existing summarizers is still as large as 0.20–0.26, suggesting that there is much room for improvement.

Results on Different Entity Types. We break down the results of existing entity summarizers into 7 entity types (i.e., classes). When $k = 5$ in Fig. 5, there is no single winner on every class, but BAFREC and MPSUM are among top three on 6 classes, showing relatively good generalizability over different entity types. Some entity summarizers have limited generalizability and they perform not well on certain classes. For example, RELIN and CD mainly rely on the self-information of a triple, while for `Location` entities their latitudes and longitudes are often unique in DBpedia but such triples with large self-information rarely appear in ground-truth summaries. Besides, most summarizers generate low-quality summaries for `Agent`, `Film`, and `Person` entities. This is not surprising since these entities are described in more triples and/or by more properties according to Fig. 2. Their summarization is inherently more difficult. When $k = 10$ in Fig. 6, MPSUM is still among top three on 6 classes. KAFCA also shows relatively good generalizability—among top three on 5 classes.

Results of Supervised Learning. As shown in Table 4, among the six supervised learning based methods, RandomForest and REPTree achieve the highest
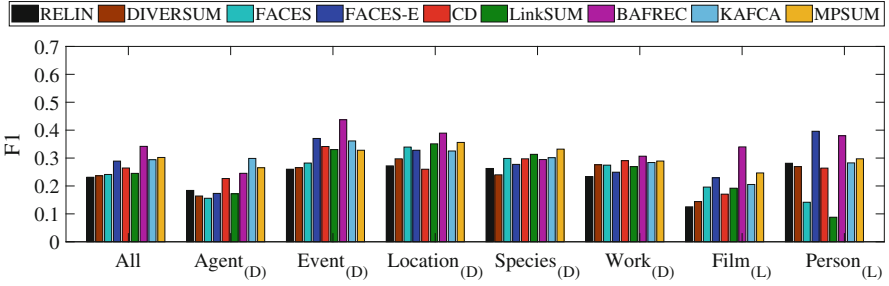
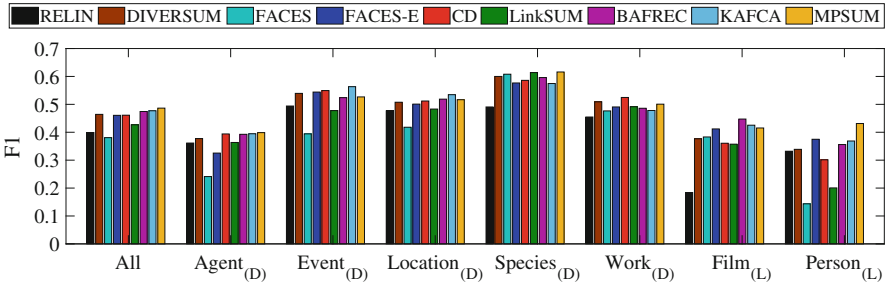**Fig. 5.** Average F1 over all the entities in each class under $k = 5$.

Fig. 6. Average F1 over all the entities in each class under $k = 10$.

F1 on DBpedia and LinkedMDB, respectively. Four methods (MultilayerPerceptron, AdditiveRegression, REPTree, and RandomForest) outperform all the existing entity summarizers on both datasets under both size constraints, and two methods (SMOreg and LinearRegression) only fail to outperform in one setting. The results demonstrate the powerfulness of supervised learning for entity summarization. Further, recall that these methods only use standard models and rely on features that are used by existing entity summarizers. It would be reasonable to predict that better results can be achieved with specialized models and more advanced features. However, creating a large number of ground-truth summaries for training is expensive, and the generalizability of supervised methods for entity summarization still needs further exploration.

Moreover, we are interested in how much the seven features contribute to the good performance of supervised learning. Table 5 shows the results of Random-Forest after removing each individual feature. Considering statistical significance at the 0.05 level, two features $\mathtt{gf}_\mathbb{T}$ and $\mathtt{lf}$ show effectiveness on both datasets under both size constraints, and two features $\mathtt{vf}_\mathbb{T}$ and $\mathtt{si}$ are only effective on LinkedMDB. The usefulness of the three binary features $\mathtt{isC}$, $\mathtt{isE}$, and $\mathtt{isL}$ is not statistically significant.

Conclusion. Among existing entity summarizers, BAFREC generally shows the best performance on ESBM while MPSUM seems more robust. However, none of them are comparable with our straightforward implementation of supervised

Table 5. F1 of RandomForest after removing each individual feature, its difference from using all features ($\Delta\%$), and the significance level for the difference (p).

DBpedia								LinkedMDB							
$k=5$				$k=10$				$k=5$				$k=10$			
	F1	$\Delta\%$	p		F1	$\Delta\%$	p		F1	$\Delta\%$	p		F1	$\Delta\%$	p
All	0.399	—	—	All	0.576	—	—	All	0.449	—	—	All	0.506	—	—
-gf$_\mathbb{T}$	0.346	−5.360	0.000	-lf	0.546	−0.030	0.000	-gf$_\mathbb{T}$	0.383	−0.066	0.000	-lf	0.473	−0.033	0.008
-lf	0.366	−3.307	0.000	-gf$_\mathbb{T}$	0.551	−0.025	0.000	-lf	0.413	−0.036	0.025	-vf$_\mathbb{T}$	0.477	−0.029	0.010
-isC	0.392	−0.720	0.261	-vf$_\mathbb{T}$	0.569	−0.007	0.198	-vf$_\mathbb{T}$	0.414	−0.035	0.022	-gf$_\mathbb{T}$	0.479	−0.027	0.007
-isE	0.397	−0.267	0.720	-isE	0.570	−0.006	0.262	-si	0.442	−0.007	0.574	-si	0.486	−0.020	0.009
-si	0.400	+0.027	0.973	-isC	0.571	−0.005	0.303	-isE	0.455	+0.005	0.651	-isL	0.491	−0.015	0.079
-isL	0.401	+0.160	0.816	-si	0.572	−0.004	0.402	-isL	0.456	+0.007	0.504	-isE	0.492	−0.014	0.148
-vf$_\mathbb{T}$	0.407	+0.720	0.346	-isL	0.578	+0.002	0.683	-isC	0.463	+0.013	0.281	-isC	0.514	+0.008	0.396

learning, which in turn is still far away from the best possible performance represented by ORACLE. Therefore, entity summarization on ESBM is a non-trivial task. We invite researchers to experiment with new ideas on ESBM.

6   Discussion and Future Work

We identify the following limitations of our work to be addressed in future work.

Evaluation Criteria. We compute F1 score in the evaluation, which is based on common triples but ignores semantic overlap between triples. A triple t in a machine-generated summary S may partially cover the information provided by some triple t' in the ground-truth summary. It may be reasonable to not completely penalize S for missing t' but give some reward for the presence of t. However, it is difficult to quantify the extent of penalization for all possible cases, particularly when multiple triples semantically overlap with each other. In future work, we will explore more proper evaluation criteria.

Representativeness of Ground Truth. The ground-truth summaries in ESBM are not supposed to represent the view of the entire user population. They are intrinsically biased towards their creators. Besides, these ground-truth summaries are created for general purposes. Accordingly, we use them to evaluate general-purpose entity summarizers. However, for a specific task, these summaries may not show optimality, and the participating systems may not represent the state of the art. Still, we believe it is valuable to evaluate general-purpose systems not only because of their wide range of applications but also because their original technical features have been reused by task-specific systems. In future work, we will extend ESBM to a larger scale, and will consider benchmarking task-specific entity summarization.

Form of Ground Truth. ESBM provides ground-truth summaries, whereas some other benchmarks offer ground-truth scores of triples [1,13,22]. Scoring-based ground truth may more comprehensively evaluate an entity summarizer than our set-based ground truth because it not only considers the triples in a

machine-generated summary but also assesses the rest of the triples. However, on the other hand, a set of top-scored triples may not equal an optimal summary because they may cover limited aspects of an entity and show redundancy. Therefore, both methods have their advantages and disadvantages. In future work, we will conduct scoring-based evaluation to compare with the current results.

References

1. Bobic, T., Waitelonis, J., Sack, H.: FRanCo - a ground truth corpus for fact ranking evaluation. In: SumPre 2015 & HSWI 2015 (2015)
2. Cheng, G., Tran, T., Qu, Y.: RELIN: relatedness and informativeness-based centrality for entity summarization. In: Aroyo, L., et al. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 114–129. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25073-6_8
3. Cheng, G., Xu, D., Qu, Y.: C3D+P: a summarization method for interactive entity resolution. J. Web Semant. **35**, 203–213 (2015). https://doi.org/10.1016/j.websem.2015.05.004
4. Cheng, G., Xu, D., Qu, Y.: Summarizing entity descriptions for effective and efficient human-centered entity linking. In: WWW 2015, pp. 184–194 (2015). https://doi.org/10.1145/2736277.2741094
5. Gottschalk, S., Demidova, E.: EventKG - the hub of event knowledge on the web - and biographical timeline generation. Semant. Web **10**(6), 1039–1070 (2019). https://doi.org/10.3233/SW-190355
6. Gunaratna, K.: Semantics-based summarization of entities in knowledge graphs. Ph.D. thesis, Wright State University (2017)
7. Gunaratna, K., Thirunarayan, K., Sheth, A.P.: FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In: AAAI 2015, pp. 116–122 (2015)
8. Gunaratna, K., Thirunarayan, K., Sheth, A., Cheng, G.: Gleaning types for literals in RDF triples with application to entity summarization. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 85–100. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-34129-3_6
9. Hasibi, F., Balog, K., Bratsberg, S.E.: Dynamic factual summaries for entity cards. SIGIR **2017**, 773–782 (2017). https://doi.org/10.1145/3077136.3080810
10. Hassanzadeh, O., Consens, M.P.: Linked movie data base. In: LDOW 2009 (2009)
11. Kim, E.K., Choi, K.S.: Entity summarization based on formal concept analysis. In: EYRE 2018 (2018)
12. Kroll, H., Nagel, D., Balke, W.T.: BAFREC: balancing frequency and rarity for entity characterization in linked open data. In: EYRE 2018 (2018)
13. Langer, P., et al.: Assigning global relevance scores to DBpedia facts. In: ICDE Workshops 2014, pp. 248–253 (2014). https://doi.org/10.1109/ICDEW.2014.6818334
14. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134

15. Liu, Q., Cheng, G., Gunaratna, K., Qu, Y.: Entity summarization: state of the art and future challenges. CoRR abs/1910.08252 (2019), http://arxiv.org/abs/1910.08252
16. Pouriyeh, S.A., Allahyari, M., Kochut, K., Cheng, G., Arabnia, H.R.: Combining word embedding and knowledge-based topic modeling for entity summarization. In: ICSC 2018, pp. 252–255 (2018). https://doi.org/10.1109/ICSC.2018.00044
17. Pouriyeh, S.A., Allahyari, M., Kochut, K., Cheng, G., Arabnia, H.R.: ES-LDA: entity summarization using knowledge-based topic modeling. In: IJCNLP 2017, vol. 1, pp. 316–325 (2017)
18. Sim, S.E., Easterbrook, S.M., Holt, R.C.: Using benchmarking to advance research: a challenge to software engineering. In: ICSE 2003, pp. 74–83 (2003). https://doi.org/10.1109/ICSE.2003.1201189
19. Stoilos, G., Stamou, G.B., Kollias, S.D.: A string metric for ontology alignment. In: ISWC 2005, pp. 624–637 (2005). https://doi.org/10.1007/11574620_45
20. Sydow, M., Pikula, M., Schenkel, R.: The notion of diversity in graphical entity summarisation on semantic knowledge graphs. J. Intell. Inf. Syst. **41**(2), 109–149 (2013). https://doi.org/10.1007/s10844-013-0239-6
21. Thalhammer, A.: Linked data entity summarization. Ph.D. thesis, Karlsruher Institut für Technologie (2017)
22. Thalhammer, A., Knuth, M., Sack, H.: Evaluating entity summarization using a game-based ground truth. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part II. LNCS, vol. 7650, pp. 350–361. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35173-0_24
23. Thalhammer, A., Lasierra, N., Rettinger, A.: LinkSUM: using link analysis to summarize entity data. In: Bozzon, A., Cudre-Maroux, P., Pautasso, C. (eds.) ICWE 2016. LNCS, vol. 9671, pp. 244–261. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-38791-8_14
24. Thalhammer, A., Rettinger, A.: Browsing DBpedia entities with summaries. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8798, pp. 511–515. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11955-7_76
25. Thalhammer, A., Toma, I., Roa-Valverde, A.J., Fensel, D.: Leveraging usage data for linked data movie entity summarization. In: USEWOD 2012 (2012)
26. Tonon, A., et al.: Contextualized ranking of entity types based on knowledge graphs. J. Web Semant. **37−38**, 170–183 (2016). https://doi.org/10.1016/j.websem.2015.12.005
27. Wei, D., Gao, S., Liu, Y., Liu, Z., Huang, L.: MPSUM: entity summarization with predicate-based matching. In: EYRE 2018 (2018)
28. Xu, D., Zheng, L., Qu, Y.: CD at ENSEC 2016: generating characteristic and diverse entity summaries. In: SumPre 2016 (2016)