Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
4(1): 1–18
DOI: 10.3934/aci.2024001
Received: 27 November 2023
Revised: 23 January 2024
Accepted: 25 January 2024
Published: 02 February 2024

Research article

Linguistic summarisation of multiple entities in RDF graphs

Elizaveta Zimina[1], **Kalervo Järvelin**[1], **Jaakko Peltonen**[1], **Aarne Ranta**[2], **Kostas Stefanidis**[1] **and Jyrki Nummenmaa**[1,*]

[1] Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

[2] Department of Computer Science and Engineering, University of Gothenburg, Sweden

* **Correspondence:** Email: jyrki.nummenmaa@tuni.fi; Tel: +358-40-52-77-999.

Academic Editor: Pasi Fränti

Abstract: Methods for producing summaries from structured data have gained interest due to the huge volume of available data in the Web. Simultaneously, there have been advances in natural language generation from Resource Description Framework (RDF) data. However, no efforts have been made to generate natural language summaries for groups of multiple RDF entities. This paper describes the first algorithm for summarising the information of a set of RDF entities in the form of human-readable text. The paper also proposes an experimental design for the evaluation of the summaries in a human task context. Experiments were carried out comparing machine-made summaries and summaries written by humans, with and without the help of machine-made summaries. We develop criteria for evaluating the content and text quality of summaries of both types, as well as a function measuring the agreement between machine-made and human-written summaries. The experiments indicated that machine-made natural language summaries can substantially help humans in writing their own textual descriptions of entity sets within a limited time.

Keywords: entity summarisation; linguistic summarisation; linked data; RDF; text generation; natural language

1. Introduction

Resource Description Framework, commonly known as RDF, is a standard defined by World Wide Web Consortium, also known as W3C, and aimed for representing linked data. According to RDF, linked data consists of subject-predicate-object triples. More specifically, an RDF triple <a,b,c >contains: a) a node for the subject, b) an arc that goes from a subject to an object for the predicate, and c) a node for the object. For example, we can represent the statement ”Stanley Kubrick directed Dr. Strangelove”, by

triple <Stanley Kubrick, directed the film, Dr. Strangelove>. An RDF graph is a collection RDF triples, where the RDF triples are edges, directed from the subject to the object.

With the increasing interest in Linked Data, RDF collections nowadays are continuously growing. For example, DBpedia's Largest Diamond Dataset describes 220 million entities by means of 1.45 billion triples*.

There are tasks where human users need to achieve an overall level of understanding of an RDF data collection. Manual processing by human users greatly complicates tasks related to analysing large quantities of such data, and the demand for automated systems facilitating these processes is increasing [1].

Summaries of RDF database contents can be utilised to aid human users. There are different ways to abstract or summarize data. For some purposes, keywords extraction [2] may be appropriate and sufficient. In particular, entity summarisation has attracted much attention in the information retrieval community. Various general purpose retrieval systems extract a small subset of the triples of a single entity, enabling convenient browsing of the entity's data [3–5]. Recently, multi-entity summarisation has been applied e.g. to entity linking [6] and entity resolution [7, 8], and for general summarisation [9, 10].

However, the summaries generated this far have not been in natural language. In this work, we generate descriptions of multiple entities in natural language and study the descriptions' usefulness.

There are situations when a user might want to read a short text describing an entity set to find out if a query retrieves a relevant collection of entities. Or, a user might need to write an abstract on some topic described in an RDF database. We suggest that in such cases an easy-to-read text helps to get an overall understanding of the set of entities.

In this work, we build an automatic natural language summarisation system for multiple entities. More specifically, we describe an algorithm extracting information from a set of entities and summarising it in natural language, and we evaluate the summaries' usefulness.

The work consisted of the following major steps: (1) developing a summarisation algorithm, (2) designing and running experiments to compare machine-made summaries, with those written by humans both with and without the help of machine-made summaries, (3) analysing the experiments.

Novel contributions. Since we have not met similar systems, we tackled some research problems at every step. *At the level of algorithm design,* our contributions are that we established a principle for extracting property-object pairs related to entities in a set to be selected for summarisation, and built a mechanism for converting the summarised structured knowledge into natural language.

At the level of experimentation, we designed an experimental approach to evaluation based on support to human task performance. We then carried out a pilot test of our machine-made summaries as opposed to human-written summaries, controlling the selection of summary topics, sizes of entity sets, timing constraints and data provided for the participants to demonstrate our system's advantages and drawbacks as clearly as possible.

At the level of analysis, we (1) established the evaluation criteria for summary quality, such as level of information aggregation and relevance, (2) defined the features of a well-constructed summary, (3) showed under which conditions machine-made summaries can outperform human-written summaries, (4) proposed a scoring function to measure the extent to which humans borrow information from machine-made summaries, if they are made available, regarded as a proxy for the level of usefulness of machine-made summaries for humans' summarisation under different conditions.

*https://www.dbpedia.org/resources/knowledge-graphs/

The experiments showed that under tight time constraints our summaries are useful to humans in writing natural language summaries on entity sets of varying length and difficulty. The shorter the time and the more difficult a topic, the more useful a machine-made summary tends to become. The levels of information aggregation and relevance, and text quality of machine-made summaries, according to our criteria, are generally higher than those of human-written summaries.

2. Previous work on entity summarisation

Significant research efforts have been devoted to entity-focused summarisation – understanding entities broadly, as a concrete or an abstract concept, such as a person, place, event, etc. Examples include textual summaries for geo-located entities and their images [11], the use of named entities in clustering news streams for news summarisation [12], and ontology-based entity recognition in multi-document summarisation [13].

For Linked Data, a number of systems provide single entity summarisation: they output k best features of one entity. These systems often employ ranking algorithms such as PageRank [3–5]. RELIN [3] elaborates a variant of a random surfer model that ranks entities' features based on their informativeness and relatedness. SUMMARUM [4] uses the popularity of resources selected by means of the corresponding Wikipedia pages. LinkSUM [5] combines the PageRank algorithm, which evaluates the importance of the connected resources, and the Backlink method, estimating the strength of the connection.

FACES [14] groups facts through hierarchical conceptual clustering to find most representative ones. DIVERSUM [15] is a diversity-based system of graphical entity summarisation. It produces a summary as a graph, but not a feature set. [16] extract unexpected facts about entities.

REMES [9] makes general purpose summaries, for several entities analysing graph- and semantics-based relatedness between entities focusing on diversity. [17] build preview tables presenting the main entity types and their relationships.

[6] developed a system for entity linking, helping humans to connect entities in text with corresponding entities in a knowledge base by summaries of entity descriptions through the analysis of characterising and differentiating power, information overlap, and relevance to context. [18] generate compact summaries of entity descriptions for entity resolution.

C3D+P [7] focuses on commonalities and conflicts between two entities. [8] uses comparative tables, outperforming FACES and C3D+P.

More and more often, entity summarisation employs neural networks [19–21], which are also involved in generating natural language summaries for single entities [22–24]. These systems are trained on the corpora of textual summaries aligned with structured data.

Most of the systems producing summaries as a list of best features are evaluated against ground-truth summaries by means of various metrics such as F1 [19, 20]. The output in the form of natural language is often evaluated with such metrics as BLEU and ROUGE also in comparison with gold standard texts [23, 24].

To our knowledge, there have been no works for natural language summaries of groups of multiple RDF entities. In our task, we operate with sets of entities and their structured features with no corresponding textual summaries. The lack of model texts raises a question on what a good summary of RDF triples should look like. Following the ideas of [3, 25, 26], we put forward several assessment

criteria.

3. Generation of linguistic machine-made summaries for multiple entities

We developed a prototype system, which, given a SPARQL query and a corresponding collection of entities[†], selects the most common features of these entities and summarises the features in natural language. The input data are used to extract (if possible) the entities' ontology class and their most common features, grouped by their corresponding properties. The extracted information is then supplemented with world knowledge by means of the ranking system and verbalised as natural text.

3.1. Analysis and verbalisation of ontology classes and properties

We consider an entity in an RDF graph as a URI (uniform resource identifier) characterised by a number of features as property-object pairs, represented as triples in the RDF graph. Entities in a set can belong to a particular ontology class or classes, which can be explicitly stated in a SPARQL query, e.g.

In which television shows has Steven Moffat served as an executive producer?
`{?uri dbo:executiveProducer res:Steven_Moffat. ?uri rdf:type dbo:TelevisionShow}`

When the ontology class is unknown, we first extract class values of entities and check if they belong to the *Person* class (for grammatical agreement), otherwise we use the word *entities*.

For every entity we obtain sets of features and so-called inverse features, represented by triples where an entity is an object, but not a subject. For example, the entity *Rome* has the feature `res:Rome dbo:country res:Italy`, while it also appears as an object: `res:Ornella_Muti dbo:birthPlace res:Rome.`

We normalise the data defining the synonymous properties within the entities. We utilise the experience of [27] in question answering over DBPedia, where properties with identical or synonymous meaning were grouped for semantically flexible search. We calculate the frequency distribution of property-object pairs across the entities. The summarisation system filters out features contained in less than 10% of entities in a set. If the summary would go below 80 words, the limit is lowered to 5%. Our experimentation indicates that 80 words are sufficient to provide general information on a topic and the experimentation also supported the choice of the percentages. No upper limit was used. In our experiments the summaries never exceeded 350 words.

To produce a natural language summary from the extracted features, we utilise the resources from [27], where verbalisations of frequent DBPedia properties were collected. We use the verb–property correlations in the opposite direction, e.g.:

```
{"birthPlace": "were born in",
 "university": "studied at",}
 "headquarter": "are headquartered in"}.
```

To form a natural language sentence, we put a subject (or it's statistical percentage) before the verb phrase, and a value after it, e.g.:

`20% of the persons were born in the USA.`

[†]In our experiments, we are using DBpedia 2016-10.

As subjects are in plural, we use *are/were* in predicates with participial and nominative constructions. With people we use past tense as in biographies (*were born, studied at*, etc.), otherwise the present tense.

However, not all facts about people are related to the past, and some properties related to non-person entities semantically require the past tense, usually for properties describing actions from the past, for example:

```
{"writer": "were written by",
 "discoverer": "were discovered by", ...}
```

For properties not in the verbalisation database, the standard construction *the [property] of ... is/was* is applied.

Some properties have several verbalisations, depending on the subject class. For example, the properties *location, country, city, region,* etc. can be expressed by *to be located in* when speaking about objects having the physical location quality (e.g. companies, rivers, museums, etc.). Other objects, such as cars, films or television shows, can also have these properties, but it would be incorrect to say that they *are located* somewhere. These special-case classes for certain properties are kept separately and invoke the universal construction *the [country/location, etc.] of ... is.*

3.2. Mentioning entities based on the world knowledge

To our mind, a descriptive and syntactically varied summary should mention meaningful and popular example entities. Presently, we use the entity ranking system of LinkSum [5].

We extract the five most important entities and list them in the beginning of a summary:

Found 335 newspapers. They include The New York Times, The Guardian, The Washington Post, The Times, and The Wall Street Journal.

Notably, five is a value found through experimentation. Some other value could have worked also.

Our system also checks all values corresponding to the properties mentioned in the summary. The ranking system selects the Top-5 with the highest rank (being not less than 50 – experimentally found value) to be mentioned in the text:

The genre of 10.30% of the record labels is Heavy metal music. Other genres include Hard rock, Progressive rock, Punk rock, Indie rock, Hardcore punk, etc.

3.3. Summary example

Below is a summary generated by our system for a DBpedia entity set on the topic *People who have won all four major annual American entertainment awards: Emmy, Grammy, Oscar, and Tony* (15 entities, 2343 triplets after data cleaning):

Found 15 persons. They include Andrew Lloyd Webber, Richard Rodgers, Whoopi Goldberg, John Gielgud, John Legend, etc. 20.00% of the persons were born in New York City. Other birth places included also: Washington, D.C., Berlin, Puerto Rico, Manhattan, etc. 60.00% of the persons were born before the 1950s, 13.33% – in the 1950s, 13.33% – in the 1970s. 53.33% of the persons were actors, 46.67% – musical artists, 40.00% – songwriters, 26.67% – writers, 20.00% – singers, 20.00% – composers. The genre of 20.00% of the persons was Musical theatre. Other genres included also: Satire, Parody, and Film score. 13.33% of the persons died in 1993.

Summaries can show that the RDF data is not always well-ordered, e.g. the above birthplaces include countries, cities and even a city borough (*Manhattan*). The system reformulated the related properties such as *birthplace, birthdate* and *deathyear* into verb phrases *were born in* and *died in*. Numerical values, such as birthdates, were grouped. The abundance of statistical figures may sound somewhat artificial, but they are informative and can help in choosing relevant features.

4. Experiment design

In the experiments we compared machine-made summaries with those written by humans both with and without the help of machine-made summaries. We explored the advantages of machine-made summaries and revealed some ways for their improvement. We also explored the helpfulness of machine-made summaries for persons writing a short text about an entity set within a limited time.

We recruited 20 people having at least a bachelor's degree to write 6 summaries, each from an entity set of some topic. The participants were recruited from the campus and they could also ask their friends. The participants were given a brief introduction to RDF and DBPedia, a description of the data provided, and guidelines for writing summaries. For the last 3 summaries, also a machine-made summary was provided. The participants were instructed that they were free to use and modify the machine-made summaries as they wished when producing summaries. The desired length of a summary was 80–350 words – generally enough to cover the topic briefly. The participants received a movie ticket as a reward for participation.

We carried out two experiments involving 10 persons each, otherwise equal, but the first group had maximum 15 minutes (hereinafter 15-minute experiment) to spend on each summary, and the second group maximum of 10 minutes (10-minute experiment). The temporal limits were determined after a study, which indicated that machine-made summaries are especially helpful for writing a summary of our required length in a time between 10 and 15 minutes.

The participants were asked to use only the information of the DBPedia pages from the lists and to aggregate that information for generalisation. We provided a summary example, but noted that a participant could choose some other strategy of writing:

Stanley Kubrick directed 16 films. They include Dr. Strangelove, 2001: A Space Odyssey, Full Metal Jacket, A Clockwork Orange, and The Shining. The films were made in the 1950s–1990s. The film scripts were written by Stanley Kubrick, Vladimir Nabokov, Howard Sackler, etc. The producers included Stanley Kubrick, James B. Harris and Jay Bonafield. Gerald Fried was among the music composers. The films were distributed by Warner Bros., United Artists, Metro-Goldwyn-Mayer, Universal Studios, Columbia Pictures, etc.

The sets of the 6 topics varied in size (Table 1). We considered a set as *small* if within the limited time a person could browse all of its entity links, *medium* if a participant could at least study the entire list of entity names, and *large* if the number of entities was too large to view all entity names. This produced the following entity set size ranges: $5 < \text{small} < 20$; $60 < \text{medium} < 120$; $200 < \text{large}$.

The first half of the experiment (without machine-made summaries) included small, medium and large sets, and the second half (with machine-made summaries) contained the other small, medium and large sets. Different participants worked on different combinations of sets with and without machine-made summaries. We obtained an equal number of summaries for every set: in total, 30 summaries written in 15 minutes with the help of machine-made summaries and 30 summaries written without

Table 1. Topics and entity set sizes.

Size	Number of Entities	Topic	Short Title
Small	12	Works by Oscar Wilde	Wilde
Small	15	Fifteen people who have won all four major annual American entertainment awards: Emmy, Grammy, Oscar and Tony (EGOT)	EGOT
Medium	88	Works by Arthur Conan Doyle	Doyle
Medium	114	People who have been appointed and confirmed as justices to the Supreme Court of the United States	SC
Large	211	All football players that have ever played in Manchester United F.C.	MU
Large	925	All Nobel laureates	Nobel

machine-made summaries, and the same numbers for the 10-minute tasks, yielding 120 human-written summaries overall.

After the task, participants filled in a questionnaire assessing on a Likert scale the summary writing difficulty and their prior knowledge on topics.

4.1. Selection of criteria for assessing the summaries

While text summarisation criteria have been studied, we found no literature on assessing natural language summaries of RDF data. A person writing a text summary can see the whole input data; in multi-entity summarisation this is not the case, especially with medium and large sets and limited time frames.

We propose general criteria for well-constructed summaries of RDF data.

– *Content aggregation quality*: To generalise, the summary should contain as much aggregated data as possible, information on common (frequent) qualities of the entities and aggregated temporal and numerical data.

– *Content relevance quality*: Relevance: the information in a summary as a whole should be relevant to the subject of writing.

– *Text quality*: A summary should be linguistically correct.

The evaluation of human-written and machine-made summaries involved two persons, giving 1 to 5 points on Likert scale to each criterion. The agreement between the evaluators was 83% for the level of aggregation, 78% for relevance, and 80% for text quality. The difference in evaluation never exceeded 1 Likert point.

4.2. Aggregation assessment criteria

With limited time to familiarise themselves with data and write a summary, humans may have difficulties choosing facts for aggregation objectively, unlike an algorithm that can output statistics on a large data set in milliseconds. Bearing this in mind, we gave aggregation points not only to summaries that seemed to contain perfectly truthful generalising data, but also to summaries merely reflecting the author's *effort* to aggregate facts collected. Without this effort a summary falls into enumeration of separate facts about individual entities and in fact cannot be called a summary.

The evaluators marked up the aggregating and non-aggregating ideas in the summaries. By an aggregating idea we understand a sentence or a part of it describing some feature/features of more than one DBpedia entities. For example, in the summary example provided in *Setup of the Experiments* all sentences are attributed to aggregating ideas, except the sentence *Gerald Fried was among the music composers.* The latter provides information on only one entity (as opposed to e.g. an aggregating idea in a statement that Gerald Fried was a composer of several of Kubrick's films). Other sentences mention properties that several entities have (*scriptwriter, producer, distributor*). The second sentence can be also attributed to aggregating ideas since it exemplifies the previous sentence, saying that all entities belong to the *film* class.

After marking the summaries, the evaluators excluded all stop words and counted the number of (content) words. The scores were based on the percentage of content words belonging to aggregating ideas: 1 – less than 25%, 2 – 25–50%, 3 – 51–65%, 4 – 66–80%, 5 – over 80%.

4.3. Relevance assessment criteria

The relevance of summaries to the topics was scored at five levels based on the percentage of content words belonging to relevant ideas: 1 – less than 25%, 2 – 25–50%, 3 – 51–65%, 4 – 66–80%, and 5 – over 80%.

We understand *relevant* as corresponding to the topic. An idea can be considered relevant if it expresses general information on the topic or a notable fact. For example, the sentence *The term of William O. Douglas, lasting 36 years and 209 days, is the longest term in the history of the Supreme Court* is a notable fact about the Supreme Court judges, describing a unique quality of one of the entities and contributing to the development of the topic.

On the other hand, random facts about some entities from a set are not similarly relevant to the topic. For example, *John Marshall was a member of Federalist Party* is not considered relevant, since there were other representatives of this party in the Supreme Court.

4.4. Text quality assessment criteria

The summaries were given the following Likert scale points in terms of linguistics when they contained: 1 – 8 or more mistakes, 2 – 6–7 mistakes, 3 – 4–5 mistakes, 4 – 2–3 mistakes, 5 – 0–1 mistakes.

A *mistake* here means a clear violation of the orthographic, morphological, syntactic and major punctuation norms. Disputable situations were resolved in favour of the summary writer. If the same type of mistake is repeated several times in text, it is counted as one.

5. Experiment results

For the statistics, we mark p-values significant at 0.01 level in **bold**. In the tables, for brevity, human-written summaries are denoted by HSs and machine-made summaries are denoted by MSs.

5.1. Text length

In the 10-minute experiment the participants reached the required 80 words length significantly more often with the help of machine-made summaries (Fisher's exact test, $p \approx \mathbf{0.0009}$); however, this

Table 2. Numbers of human-written summaries of less than the recommended minimum of 80 words (out of 30 summaries in each group).

Experiment	MSs available	HSs of less than 80 words
10-minute experiment	No	17
	Yes	4
15-minute experiment	No	4
	Yes	2

was not the case in the 15-minute experiment ($p \approx 0.6707$; Table 2). In the 10-minute experiment, human summaries written with the help of machine-made summaries were significantly longer than those without the help of machine-made summaries (Mann-Whitney U-test over all HSs of all topics, $p \approx \mathbf{0.000097}$; Table 3). Notably, the machine-made summaries were on the average longer than the human-written summaries.

Table 3. Average lengths of HSs and MSs (words).

Topic	HSs in 10-Minute Experiment		HSs in 15-Minute Experiment		MSs
	Without MSs	With MSs	Without MSs	With MSs	
Wilde	84.8	95.8	97.8	125.8	130.0
EGOT	68.2	97.4	104.2	90.8	129.0
Doyle	80.0	123.6	106.0	84.8	146.0
SC	82.4	124.6	89.0	147.4	310.0
MU	75.6	102.0	93.8	114.4	201.0
Nobel	78.6	130.0	94.0	91.2	157.0
Average	78.3	**112.2**	97.5	109.1	178.8

Sometimes the participants were clearly in trouble finding enough relevant information. Enumerating separate random facts about entities instead merely decreased the summary content quality. The participants more often reached the word minimum when they had machine-made summaries at hand. Obviously, having some supportive natural language text together with the raw data facilitated summary writing.

5.2. Content quality

In the 10-minute experiment, summaries written with machine-made summaries available were on average of significantly higher quality than those without machine-made summaries for both Aggregation and Relevance (Mann-Whitney U-test over all topics, $p \approx \mathbf{0.0001}$ for Aggregation, $p \approx \mathbf{0.0085}$ for Relevance; Table 4).

In the 15-minute experiment, average differences were not statistically significant (Mann-Whitney U-test over all topics, $p \approx \mathbf{0.1344}$ for Aggregation, $p \approx \mathbf{0.0862}$ for Relevance; Table 5).

The Likert scale scores were given regardless of the summary length. Some short and not very comprehensive summaries got high scores for aggregation and relevance. However, we consider a summary well-constructed if it is at least 80 words long and got 4 or 5 for both aggregation and relevance

Table 4. Content quality of human-written summaries written in 10 minutes (Mean ± Standard Deviation).

Topic	HSs without MS		HSs with MS	
	Aggregation	**Relevance**	**Aggregation**	**Relevance**
Wilde	2.8±1.0	4.4±0.8	4.0±1.6	4.2±1.6
EGOT	2.4±1.7	3.0±1.7	3.8±1.1	4.2±0.8
Doyle	2.8±1.5	3.6±1.7	4.8±0.4	5.0±0.0
SC	2.4±1.4	3.8±1.1	4.2±1.6	4.2±1.6
MU	2.6±1.2	2.4±1.4	3.8±1.6	3.8±1.6
Nobel	3.0±1.8	3.2±1.8	4.0±1.3	4.6±0.8
Average	2.7±1.4	3.4±1.4	**4.1±1.3**	**4.3±1.1**

Table 5. Content quality of human-written summaries written in 15 minutes (Mean ± Standard Deviation).

Topic	HSs without MS		HSs with MS	
	Aggregation	**Relevance**	**Aggregation**	**Relevance**
Wilde	3.0±1.3	4.4±1.2	3.0±1.7	3.6±1.7
EGOT	4.4±0.8	3.4±1.6	4.6±0.5	4.8±0.4
Doyle	4.0±1.1	3.8±1.5	4.4±0.5	4.6±0.8
SC	3.2±1.5	3.8±1.5	3.4±2.0	3.4±2.0
MU	4.0±1.1	3.6±1.0	3.6±1.5	3.8±1.6
Nobel	3.2±1.5	3.4±2.0	4.4±0.8	5.0±0.0
Average	3.6±1.2	3.7±1.4	3.9±1.2	4.2±1.1

criteria, so that the content quality criteria are balanced with the length recommendation.

In the 10-minute experiment the summaries were significantly more often well-constructed when machine-made summaries were available (Fisher's exact test, $p \approx \mathbf{0.000003}$); but not so clearly in the 15-minute experiment ($p \approx 0.038$; Table 6).

Table 6. Numbers of well-constructed human-written summaries (out of 30 summaries in each group).

Experiment	MSs available	Well-constructed HSs
10-minute experiment	No	3
	Yes	21
15-minute experiment	No	11
	Yes	20

The machine-made summaries were assessed with the same content quality criteria and got 5 points for each of them, as the same criteria were used to design the system.

Admitting that the quality of summaries is person-dependent, especially within a small group of participants, we list the following observations:

1. For most of the participants, without the help of machine-made summaries, 10 minutes were not enough to write a relatively well-constructed summary. 15 minutes were still enough for only 1/3 of the summary writers.

2. Often when a summary writer was running out of time or could not find common entity information, she started picking random pieces of information to reach the recommended limit of words.

3. The machine-made summaries were of substantial help in summary writing in both experiments.

4. Content quality of the summaries was slightly higher in the 10-minute experiment with machine-made summaries than in the 15-minute experiment with machine-made summaries. The contrast of content quality with machine-made summaries versus without machine-made summaries was greater in the 10-minute experiment than in the 15-minute experiment. These observations can be explained by the higher degree of borrowing from machine-made summaries when the time is extremely limited.

5.3. Text quality

Only 2 out of 20 summary writers got 5 as linguistics score for all their texts. In most cases, the participants made 1–3 mistakes of different kinds in each summary (Table 7). The scores are higher when a machine-made summary was available, with Mann-Whitney U-test over all topics $p \approx 0.027$ for the 10-minute experiment, and $p \approx 0.3158$ for the 15-minute experiment. Machine-made summaries contained no linguistic mistakes.

Table 7. Text quality of human-written summaries (Mean ± Standard Deviation).

Topic	**HSs in 10-Minute Experiment**		**HSs in 15-Minute Experiment**	
	Without MS	**With MS**	**Without MS**	**With MS**
Wilde	5.0±0.0	4.4±0.5	4.4±0.8	3.2±0.4
EGOT	4.4±0.5	4.6±0.8	3.4±0.5	4.2±0.8
Doyle	4.8±0.4	4.6±0.5	3.0±0.0	4.4±0.5
SC	4.6±0.5	4.2±1.0	4.2±1.0	3.2±0.4
MU	5.0±0.0	4.2±0.8	4.0±0.0	3.6±0.5
Nobel	4.2±0.8	4.2±0.8	3.0±0.0	4.0±0.0
Average	4.7±0.4	4.4±0.7	3.7±0.4	3.8±0.4

There was no clear dependence between the participants' text quality and the time given. It seems to be purely dependent on a person's skills. Automatic text generation seems more likely to produce text with fewer mistakes, as even a highly literate person can easily make typos.

5.4. Keyword agreement

For better understanding of how much the summaries of participants on the same topic had in common between themselves and with the machine-made summaries, we calculated the average asymmetric overlap of keywords between them, i.e. their agreement (Tables 8 and 9).

By keywords we understand all tokens excluding punctuation, stop words and numbers. Morphological forms and derivationally related words are taken as equal keywords.

The average asymmetric agreement of two human-written summaries on a given topic is counted in the standard way (e.g. [28]) as follows. Let S_i, S_j be keyword sets of two summary texts. Their average

Table 8. Agreement and average difficulty in the 10-minute experiment (AD – average difficulty, ABP – agreement between participants, AWMS – agreement with MS).

Topic	Without MSs			With MSs		
	AD	**ABP**	**AWMS**	**AD**	**ABP**	**AWMS**
Wilde	3.4	0.2964	0.3417	2.6	0.2731	0.4218
EGOT	4.6	0.3460	0.1906	3.2	0.3852	0.3616
Doyle	3.8	0.3387	0.3509	2.8	0.5115	0.5795
SC	5.0	0.3229	0.2100	4.0	0.3429	0.4840
MU	4.2	0.2039	0.1759	3.0	0.3564	0.5511
Nobel	4.0	0.1842	0.1419	3.0	0.4181	0.5778
Average	4.1	0.2820	0.2352	3.1	0.3812	**0.4960**

Table 9. Agreement and average difficulty in the 15-minute experiment (AD – average difficulty, ABP – agreement between participants, AWMS – agreement with MS).

Topic	Without MSs			With MSs		
	AD	**ABP**	**AWMS**	**AD**	**ABP**	**AWMS**
Wilde	3.00	0.3380	0.2852	4.00	0.2523	0.3811
EGOT	2.80	0.1264	0.1423	3.25	0.2563	0.3175
Doyle	2.60	0.2307	0.3072	3.50	0.2741	0.3391
SC	5.00	0.2112	0.1919	4.20	0.2042	0.4314
MU	2.75	0.1843	0.1569	3.20	0.2501	0.4226
Nobel	3.40	0.1522	0.1944	3.00	0.3025	0.3872
Average	3.26	0.2071	0.2130	3.52	0.2566	**0.3798**

asymmetric pairwise agreement is:

$$avg\text{-}agr(S_i, S_j) = 0.5 * (|S_i \cap S_j|/|S_i| + |S_i \cap S_j|/|S_j|).$$

Through averaging, this basic formula is applied in the agreement computation of (1) all pairs of a set of human-written summaries on a topic, (2) human-written and machine-made summaries on a topic, and (3) the cross-topic average for each experiment.

The agreement between human-written summaries and the corresponding machine-made summaries were on average statistically significantly better in the setting where humans were provided with machine-made summaries than in the setting where humans were not provided with machine-made summaries (Mann-Whitney U-test over all topics, $\mathbf{p \approx 0.00000016}$ for 10-minute experiment, $\mathbf{p \approx 0.00021}$ for 15-minute experiment).

The agreement of the human-written summaries between themselves and with the machine-made summaries gets generally higher, if the participants are provided with the machine-made summaries. Seemingly, a summary writer tends to agree with at least some part of the given machine-made summary and uses it.

The shorter the time for writing, the more a participant tends to borrow from the machine-made summary. The agreement of the 10-minute summaries between themselves and with the machine-made

summaries is substantially higher almost for all topics than for those of the 15-minute experiment. A participant having more time may be more independent and creative in writing.

The shorter the time for writing, the more difficult a topic seemed for a participant without a machine-made summary and (surprisingly) the less difficult – with machine-made summary. Probably, with the machine-made summaries, the participants felt more comfortable having a (semi-)ready text from which they could use the information. As a guess, it could also be that the shorter time was somehow "just right" for the task, helping to keep the focus and concentration.

6. Discussion

Our approach is the first attempt of automatic generation of natural language summaries of multi-entity sets. We chose a statistics-based method, since we believe that the more common a feature among entities is, the more important it is and the more likely it should be mentioned in a summary. At the same time, the algorithm takes care of synonymous features that should be merged and semantically irrelevant properties that should be excluded.

For substantial evaluation of summaries, we rate information aggregation and relevance. We tried to formulate the scoring principles as strictly as possible, however evaluation is still human.

While reading the human-written summaries, we noticed outright untruthful facts extremely rarely. However, some summaries seemed more thorough and informative than others, not always depending on their length. It seems difficult to formulate precisely the principles of assessing truthfulness and comprehensiveness of summaries of RDF data. In our evaluation, we aimed to mark out obviously improper strategies of writing a summary, such as writing too short a summary or picking random facts about entity members.

The experiments showed that the task of writing a well-constructed (according to our criteria) summary is challenging, especially having very limited time and no support of a machine-made summary. The scores obtained by means of our function measuring the agreement between human-written summaries and machine-made summaries and also the retrospective questioning of the participants showed that the machine-made summaries were of substantial help, especially in the 10-minute experiment, and in many cases phrases and whole sentences from the machine-made summaries seemed to the participants good enough to use.

The tendency to copy the machine-made summary to a greater level is more pronounced in cases when a summary writer does not know the topic well. This was also the only clear conclusion we could make out of the prior knowledge of the participants.

At the same time, the participants tended to modify the texts so that they would sound less "mechanistic" and more natural. The participants often considered the machine-made summary too statistical, containing detailed numerical data never met in typical human-written texts.

The system can also help in finding inconsistencies in a data set. A database moderator can obtain a list of entities with faulty features and fix them.

Our method can be extrapolated to other knowledge bases by means of the available mappings between their structures (for example, [29, 30]). We are also planning to extend our database of lexical verbalisations of DBpedia/Wikidata properties through further work on entity/relation linking and question answering involving more extended datasets such as LC-QuAD 2.0 [31].

7. Conclusions

We have studied a previously underinvestigated issue of natural language summarisation of multiple RDF entities and implemented a system to solve this task. The experiments revealed the considerable benefit of the machine-made summaries as an easy-to-read source of aggregate information. A machine-made summary can be obtained based on practically any collection of entities, even when they are united by some topic that is not widely covered in the web. This summarised information can be obtained instantly, whereas a human needs more time to look through an entity set and decide what to write about it. Even though in most cases machine-generated text still falls behind high-quality human-written text in terms of style and content, our system successfully solves a number of application tasks and outranks average human performance. In the future, we aim to develop the algorithm into a system comparing several sets of entities and producing natural language reports on their similarities and differences in several languages by using the techniques introduced in [32] and [33] for multilingual verbalisation.

Use of AI tools declaration

The authors hereby declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research has been funded in part by the Academy of Finland, decision number 352441.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. V. Christophides, V. Efthymiou, K. Stefanidis, *Entity resolution in the Web of data*, Synthesis lectures on the Semantic Web: theory and technology, Morgan & Claypool Publishers, 2015. https://doi.org/10.1007/978-3-031-79468-1
2. H. Shah, P. Fränti, Combining statistical, structural, and linguistic features for keyword extraction from web pages, *Applied computing and intelligence*, **2** (2022), 115–132. https://doi.org/10.3934/aci.2022007
3. G. Cheng, T. Tran, Y. Qu, RELIN: relatedness and informativeness-based centrality for entity summarization, The Semantic Web–ISWC 2011, *The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10*, (2011), 114–129. https://doi.org/10.1007/978-3-642-25073-6_8
4. A. Thalhammer, A. Rettinger, Browsing DBPedia entities with summaries, *The Semantic Web: ESWC 2014 Satellite Events*, (2014), 511–515. https://doi.org/10.1007/978-3-319-11955-7_76

5. A. Thalhammer, N. Lasierra, A. Rettinger, LinkSUM: using link analysis to summarize entity data, *International Conference on Web Engineering*, (2016), 244–261. https://doi.org/10.1007/978-3-319-38791-8_14

6. G. Cheng, D. Xu, Y. Qu, Summarizing entity descriptions for effective and efficient human-centered entity linking, *Proceedings of the 24th International Conference on World Wide Web*, (2015), 184–194. https://doi.org/10.1145/2736277.2741094

7. G. Cheng, D. Xu, Y. Qu, C3d+ p: a summarization method for interactive entity resolution, *Web Semantics: Science, Services and Agents on the World Wide Web*, **35** (2015), 203–213. https://doi.org/10.1016/j.websem.2015.05.004

8. J. Huang, W. Hu, H. Li, Y. Qu, Automated comparative table generation for facilitating human intervention in multi-entity resolution, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, (2018), 585–594.

9. K. Gunaratna, A. H. Yazdavar, K. Thirunarayan, A. Sheth, G. Cheng, Relatedness-based multi-entity summarization, *Proceedings of the Twenty-national Joint Conference on Artificial Intelligence*, (2017), 1060–1066. https://doi.org/10.24963/ijcai.2017/147

10. G. Troullinou, H. Kondylakis, K. Stefanidis, D. Plexousakis, Exploring RDFS KBs using summaries, *The Semantic Web – ISWC*, (2018), 268–284. https://doi.org/10.1007/978-3-030-00671-6_16

11. A. Aker, R. Gaizauskas, Generating descriptive multi-document summaries of geo-located entities using entity type models, *J. Assoc. Inf. Sci. Tech.*, **66** (2015), 721–738. https://doi.org/10.1002/asi.23211

12. H.Chen, J. Kuo, S. Huang, C. Lin, H. Wung, A summarization system for Chinese news from multiple sources, *J. Am. Soc. Inf. Sci. Tech.*, **54** (2003), 1224–1236. https://doi.org/10.1002/asi.10315

13. E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, S. Shah, Multi-document summarization based on the Yago ontology, *Expert Syst. Appl.* **40** (2013), 6976–6984. https://doi.org/10.1016/j.eswa.2013.06.047

14. K. Gunaratna, K. Thirunarayan, A. Sheth, FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, (2015), 116–122. https://doi.org/10.1609/aaai.v29i1.9180

15. M. Sydow, M. Pikuła, R. Schenkel, The notion of diversity in graphical entity summarisation on semantic knowledge graphs, *J. Intell. Inf. Syst.*, **41** (2013), 109–149. https://doi.org/10.1007/s10844-013-0239-6

16. B. Schäfer, P. Ristoski, H. Paulheim, What is special about Bethlehem, Pennsylvania? Identifying unusual facts about DBpedia entities, *Proceedings of the ISWC 2015 Posters & Demonstrations Track*, 2015.

17. N. Yan, S. Hasani, A. Asudeh, C. Li, Generating preview tables for entity graphs, *Proceedings of the 2016 International Conference on Management of Data*, (2016), 1797–1811. https://doi.org/10.1145/2882903.2915221

18. D. Xu, G. Cheng, Y. Qu, Facilitating human intervention in coreference resolution with comparative entity summaries, *The Semantic Web: Trends and Challenges, ESWC 2014, Lecture Notes in Computer Science*, (2014), 535–549. https://doi.org/10.1007/978-3-319-07443-6_36

19. D. Wei, Y. Liu, F. Zhu, L. Zang, W. Zhou, J. Han, et al., ESA: Entity Summarization with Attention, *arXiv preprint arXiv:1905.10625*, 2019.

20. Q. Liu, G. Cheng, Y. Qu, DeepLENS: Deep Learning for Entity Summarization, *arXiv preprint arXiv:2003.03736*, 2020.

21. Q. Liu, Y. Chen, G. Cheng, E. Kharlamov, J. Li, Y. Qu, Entity Summarization with User Feedback, *ESWC 2020: The Semantic Web*, (2020), 376–392. https://doi.org/10.1007/978-3-030-49461-2_22

22. A. Chisholm, W. Radford, B. Hachey, Learning to generate one-sentence biographies from Wikidata, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, (2017), 633–642. https://doi.org/10.18653/v1/E17-1060

23. R. Lebret, D. Grangier, M. Auli, Neural Text Generation from Structured Data with Application to the Biography Domain, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (2016), 1203–1213. https://doi.org/10.18653/v1/D16-1128

24. P. Vougiouklis, H. Elsahar, L. Kaffee, C. Gravier, F. Laforest, J. Hare, et al., Neural Wikipedian: Generating Textual Summaries from Knowledge Base Triples, *Journal of Web Semantics*, **52** (2018), 1–15. https://doi.org/10.1016/j.websem.2018.07.002

25. C. Jumel, A. Louis, J. C. K. Cheung, TESA: A Task in Entity Semantic Aggregation for Abstractive Summarization, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, (2020), 8031–8050. https://doi.org/10.18653/v1/2020.emnlp-main.646

26. A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, D. Radev, SummEval: Re-evaluating Summarization Evaluation, *Transactions of the Association for Computational Linguistics*, **9** (2021), 391–409. https://doi.org/10.1162/tacl_a_00373

27. E. Zimina, J. Nummenmaa, K. Järvelin, J. Peltonen, K. Stefanidis, H. Hyyrö, GQA: grammatical question answering for RDF data, *Semantic Web Challenges: 5th SemWebEval Challenge at ESWC*, (2018), 82–97. https://doi.org/10.1007/978-3-030-00072-1_8

28. T. Saracevic, Measuring the degree of agreement between searchers, *Proceedings of the 47th Annual Meeting of the American Society for Information Science*, **21** (1984), 227–230.

29. M. Azmy, P. Shi, I. Ilyas, J. Lin, Farewell Freebase: Migrating the SimpleQuestions Dataset to DBpedia, *Proceedings of the 27th international conference on computational linguistics* (2018), 2093–2103.

30. T. Tanon, D. Vrandečić, S. Schaffert, T. Steiner, L. Pintscher, From Freebase to Wikidata: The Great Migration, *Proceedings of the 25th International Conference on World Wide Web*, (2016), 1419–1428.

31. M. Dubey, D. Banerjee, A. Abdelkawi, J. Lehmann, LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia, *International Semantic Web Conference*, (2019), 69–78. https://doi.org/10.1007/978-3-030-30796-7_5

32. M. Damova, D. Dannélls, R. Enache, M. Mateva, A. Ranta, Multilingual Natural Language Interaction with Semantic Web Knowledge Bases and Linked Open Data, in *Towards the Multilingual Semantic Web: Principles, Methods and Applications*, Buitelaar, P., Cimiano, P., Eds., Springer Berlin Heidelberg, (2014), 211–226. https://doi.org/10.1007/978-3-662-43585-4_13

33. D. Dannélls, Multilingual text generation from structured formal representations. PhD Thesis. University of Gothenburg, 2012.

Appendix

A. Instructions for the user experiment

The task is devoted to summarisation of multiple entities from DBpedia – a database with structured content from the information created in various Wikimedia projects.

A DBpedia entity is a URI that is characterised by a number of features, that is, property-object pairs. For example, click here: `http://dbpedia.org/resource/Bob_Marley`. This page is devoted to the entity Bob Marley and displays the information in tabular format: the left column contains the names of properties, the right one – the objects, in practice normally values. Most of the values are other entities and are clickable – you can also explore them a little bit to feel DBpedia's design concept better. You will work with sets of entities united by a common topic. The topic will be given in the first line of each file (e.g. The films directed by Stanley Kubrick). The following structure of the files depends on the task.

1. For files 1.txt, 2.txt, 3.txt:

You are given a list of links to DBPedia entities (starting from the 3rd line). Look through the pages of several entities and try to write a summary about them (80–350 words). Use only the information of the DBPedia pages from the list. Do not worry if the number of links is too high, try to generalise the information as much as possible. You can also write about any separate facts that you find interesting. For example, your summary can (but not must) look like the following:

Stanley Kubrick directed 16 films. They include Dr. Strangelove, 2001: A Space Odyssey, Full Metal Jacket, A Clockwork Orange, The Shining, etc. The film scripts were written by Stanley Kubrick, Vladimir Nabokov, Howard Sackler, etc. The producers included Stanley Kubrick, James B. Harris and Jay Bonafield. Gerald Fried was among the music composers. The films were distributed by Warner Bros., by United Artists, Metro-Goldwyn-Mayer, Universal Studios, Columbia Pictures, etc.

2. For files 4.txt, 5.txt, 6.txt:

You are given a list of links to DBPedia entities (starting from the 5th line) and a machine-produced summary about them (3rd line). Using the information from both sources, write your own summary (80–350 words). You are free to modify and borrow sentences from the machine-produced summary as much as you want.

For both tasks: Spend not more than 10 minutes on working with each entity set. This time should include both studying the information about the entities and writing the summary. If you cannot reach the required minimum of words, submit what you have managed to write (but still try your best). Please strictly observe the 10-minutes limit! We want to see how much a person can do exactly within this time. You can make breaks between writing summaries (but not within 1 summary). Place your texts in the beginning of the corresponding input files.

After you have done all the tasks, fill in the Follow-up questions.docx form.

Note: If you find any mistake or inconsistency in the DBpedia data or machine-generated summaries, it should not confuse you. DBpedia is a crowd-sourced community effort. Your summaries should include only facts that seem truthful to you.

AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)