Verbalizing the Evolution of Knowledge Graphs with Formal Concept Analysis

Martín Arispe Riveros[1], Mayesha Tasnim[2], Damien Graux[3], Fabrizio Orlandi[3], Diego Collarana[2](✉)

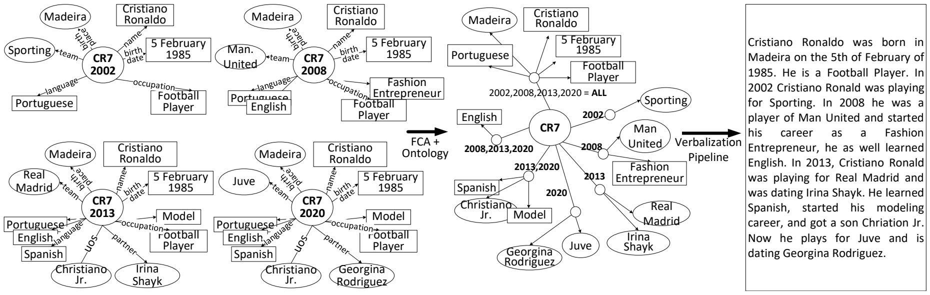
[1] Universidad Privada Boliviana, Bolivia
[2] Fraunhofer IAIS and University of Bonn, Germany
[3] ADAPT SFI Centre, Trinity College Dublin, Ireland
{diego.collarana.vargas|mayesha.tasnim}@iais.fraunhofer.de,
martinarispe@upb.edu, {orlandif|grauxd}@tcd.ie

Abstract. Questioning Answering and Verbalization over Knowledge Graphs (KGs) are gaining momentum as they provide natural interfaces to knowledge harvested from a myriad of data sources. KGs are dynamic: new facts are added and removed over time, producing multiple versions, each representing a knowledge snapshot of a point in time. Verbalizing a report of the evolution of entities is useful in many scenarios, e.g., reporting digital twins' evolution in manufacturing or healthcare. We envision a method to verbalize a graph summary capturing the temporal evolution of entities across different KG versions. Technically, our approach considers revisions of a graph over time and converts them into RDF molecules. Formal Concept Analysis is then performed on these RDF molecules to synthesize summary information. Finally, a verbalization pipeline generates a report in natural language.

1 Introduction

Talking Knowledge Graphs in the form of Question Answering and Story Telling components have gained momentum as natural user interfaces to heterogeneous data structures. On the one hand, Question Answering (QA) technology, including QAnswer [3] and WDAqua-core1 [4] paved the way to knowledge graph agnostic QA systems. On the other hand, the ability to verbalize part of a knowledge graph (KG) to create reports (storytelling), is considered another critical application in many domains, e.g., healthcare and finance. Diverse approaches have been proposed to verbalize semi- and fully-structured data. The most recent approaches focus on data-hungry deep learning architectures, e.g., [1,12]. However, these approaches perform poorly on unseen domains. Although much attention has been put on QA and Verbalization over KGs, few have considered the dynamic nature of KGs. Knowledge graphs are becoming increasingly dynamic, and approaches have been proposed [19] to (i) detect changes during their evolution, (ii) represent change information (using vocabularies) [16], and (iii) propagate changes to replicas or federated systems [5]. In this context, change detection is typically performed computing "deltas" (or changesets[1]) between two versions

[1] https://www.w3.org/2009/12/rdf-ws/papers/ws07 (accessed on 08/09/2020)



Cristiano Ronaldo was born in Madeira on the 5th of February of 1985. He is a Football Player. In 2002 Cristiano Ronald was playing for Sporting. In 2008 he was a player of Man United and started his career as a Fashion Entrepreneur, he as well learned English. In 2013, Cristiano Ronald was playing for Real Madrid and was dating Irina Shayk. He learned Spanish, started his modeling career, and got a son Chriation Jr. Now he plays for Juve and is dating Georgina Rodriguez.

Fig. 1: **Motivating example:** Four-yearly RDF entities, describing the professional evolution of the football player Cristiano Ronaldo, pass through an entity-evolution summary creation step. Finally, a verbalization step produces an understandable human summary of the entity's evolution.

of a knowledge graph at different granularity levels [19]: dataset, resource or statement. Let us consider four different entity descriptions depicted in Figure 1. These graphs represent four different years of the football player Cristiano Ronaldo, aka., CR7. Through time, CR7 evolved in different contexts of his professional career and personal life. He changed teams several times; thus, the relation `team` changed as well. He also became an entrepreneur in 2008 and a father in 2013. Developing an approach to verbalize the evolution of entities in a knowledge graph would be useful to encompass in a glance CR7's life.

More generally, creating a story (report) for KG evolution is useful in many domains and challenging as it should be schema-agnostic. For example, 1) in industry 4.0 to report the evolution of digital twins, or 2) in healthcare to understand the evolution of patients according to their records, or 3) to produce financial reports of the evolution of companies and industries. In this study, we present the vision of an approach to produce a summary of the evolution of entities in knowledge graphs. The approach is based on Formal Concept Analysis to automatically create entity summary in a schema-agnostic manner. Finally, we employ a template-based approach to verbalize the evolution. This paper is structured as follows. Section 2 presents the state of the art. Then, Section 3 defines the approach and techniques envisioned. Finally, Section 4 wraps up and outlines future work.

2 Related Work

The associated literature is threefold. First, we review verbalization efforts on structured data and KGs. Then, we present applications and benefits of FCA on KGs. Finally, we provide an overview of the current techniques developed to manage KG evolution.

KGs Verbalization. A variety of works have proposed methods for verbalizing structured data and KGs. We start with the early (but widely used) heuristic-driven methods where the main objective is to choose the right set of rules or templates to verbalize

KGs. One of the most representative works here is SimpleNLG [10] and its different variants, French, Spanish, German, and Italian. SimpleNLG defines a three-stage pipeline for Natural Language Generation (NLG). Hence, these approaches are hard to adapt to different domains requiring many tunning efforts, so recent NLG approaches employ neural network architectures. These neural approaches mostly use the seq2seq architectures with attention mechanisms [1,12] and replace the three-phase pipeline with an end-to-end approach. However, these approaches required large training data, and tend to perform poorly on unseen domains.

FCA and KGs. Recent work in the literature aimed at applying FCA to knowledge graphs for concept analysis. For example, in [9] the authors propose an extension of FCA where a dataset is a hypergraph instead of a binary table. Motivated by the fact that, thus far, FCA has been successfully applied to discover conceptual structures in tabular/relational data. Similarly, [14] proposes the Relational Concept Analysis (RCA), where FCA is adapted to graphs and applied to individual entities of different types singularly. The only relevant work applying FCA to KGs in order to analyse their evolution is presented in [11] and [18]. In both cases FCA is leveraged for the identification of differences/similarities between different versions of a KG. In this paper we extend that work by applying it to KG verbalization/summarization tasks.

KG Evolution. The increasingly dynamic nature of KGs has driven researchers into investigating solutions for managing their evolution [7]. Approaches have been proposed to: (i) *detect* changes during their evolution [19], (ii) *represent* their changes and dynamics using ontologies [16] (iii) *archive* their history [6], (iv) *propagate* their changes over federated systems [5]. Approaches for change detection mainly focus on computing "deltas" (or changesets between different versions of a KG at various granularity levels [19]: dataset, resource and statement level. The Changeset Vocabulary[2] defines a set of terms for describing changes on a resource and statement level. The DELTA-LD framework detects changes between two versions of a KG and represents them using a specific ontology [16]. In [15], in order to study the dynamics of LOD, the authors propose a framework for extracting and analysing the evolution history of LOD datasets. A commonality of all these approaches is that specific SPARQL queries need to be constructed in order to extract the changes, and the history, of a particular resource over time. This is because they all use specific ontologies to model this information. In contrast, our approach allows for automatic extraction and exploration of all changes of a class/entity over time, in an easy and accessible way.

3 Proposed Approach

Given different deltas of a knowledge graph, e.g., 2002, 2008, 2013, and 2020, and an entity type, e.g., Person. Our approach automatically produces a summary of evolution over time of the entities under the specified type. Each entity summary is comprised of the evolution of properties and relations among these entities along a temporal dimension. Finally, a verbalization step produces a summary in natural language of all

[2] `http://vocab.org/changeset/schema` (accessed on 04/09/2020)

the changes identified among the knowledge graph deltas. To better understand our approach, we define the central concepts it employs, i.e., RDF Molecule, Formal Concept Analysis, Evolution Summary, and Entity Verbalization.

3.1 Preliminaries

Definition 1 (RDF Molecule [8]). *If G is a given RDF Graph, we define an RDF Molecule M as a sub-graph of G such that,*

$$M = \{t_1, \dots, t_n\}, \forall (i,j) \in \{1, \dots, n\}^2 \, (subject(t_i) = subject(t_j))$$

where $t_1, \dots, t_n$ denote the triples in M. An RDF Molecule M consists of triples having the same subject. In this work, molecules are used as units to produce entity summaries.

Definition 2 (Formal Concept Analysis [20]). *is an algorithm aiming at grouping objects based on the overlap between their attributes. In our approach, we apply an algorithm proposed by V. Vychodil [20], after transforming RDF molecules into the binary data table it requires. Formal concepts are defined as conceptual clusters found within entity-property data tables. These data tables have rows corresponding to entities, and columns corresponding to the properties of those entities. Formal concepts are a set of $< A, B >$ pairs where A is the entity set, B is the property set, and all the entities in A contain all the properties in B. A is known as* extent *and B is known as* intent*.*

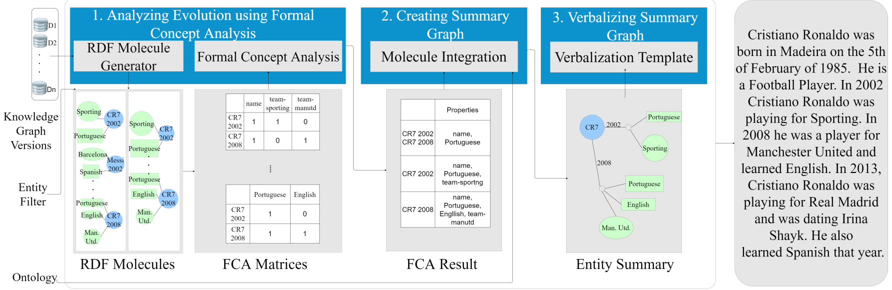
Definition 3 (Evolution Summary [18]). *To produce a temporal evolution summary of entities spread over different versions of a knowledge graph, we resort to the concept of fusion policies defined by Collarana et al. [2]. A fusion policy is a set of rules operating on the triple level, which are triggered by a certain combination of predicates and objects. Fusion policies resort to an ontology O to resolve possible conflicts and inequalities on the levels of resources, predicates, objects and literals.*

Definition 4 (Verbalization Function). *We define a verbalization function $V(t)$ which takes as input a set of triples $t = \{t_1, \dots, t_n\}$ and produces S_{Eng}, a text description of the given triples in natural (English) language.*

3.2 Architecture

Based on the summarization technique proposed by Tasnim et al. [17,18], we propose a pipeline capable of automatically verbalizing the evolution of RDF entities. Thus providing a solution to the problem of generating natural language reports on the temporal evolution of entities over different versions of a Knowledge Graph. We propose a three-fold approach, namely: identifying equivalent entities in different versions of a knowledge graph, summarizing the temporal evolution of these entities, and finally verbalizing the obtained summary.

Figure 2 depicts the main components of our architecture. First, the pipeline takes as input a set of knowledge graphs; each graph represents different time version of the same knowledge graph. These graphs are then turned into a set of RDF molecules representing groups of equivalent entities, i.e., different temporal versions of the same



	name	team-sporting	team-manutd
CR7 2002	1	1	0
CR7 2008	1	0	1

⋮

	Portuguese	English
CR7 2002	1	0
CR7 2008	1	1

	Properties
CR7 2002 CR7 2008	name, Portuguese
CR7 2002	name, Portuguese, team-sportng
CR7 2008	name, Portuguese, Engllish, team-manutd

Cristiano Ronaldo was born in Madeira on the 5th of February of 1985. He is a Football Player. In 2002 Cristiano Ronaldo was playing for Sporting. In 2008 he was a player for Manchester United and learned English. In 2013, Cristiano Ronaldo was playing for Real Madrid and was dating Irina Shayk. He also learned Spanish that year.

Fig. 2: **Architecture.**

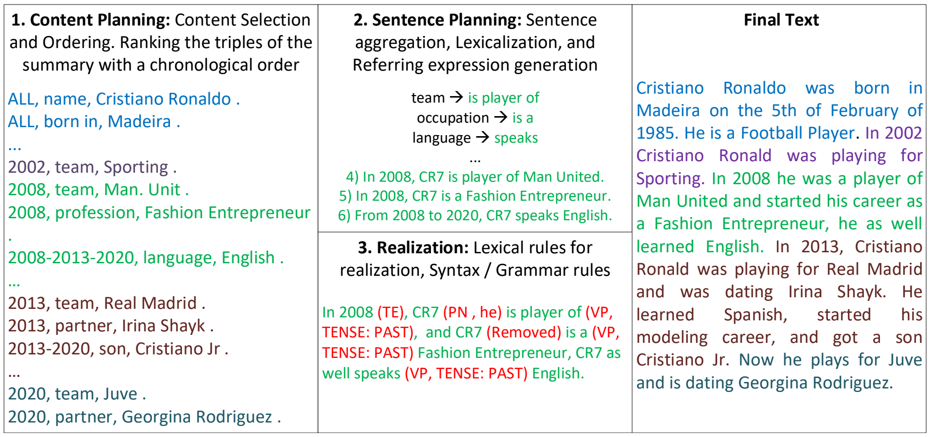
real-world entity (CR7 for example). Each group of equivalent molecules is then converted into a binary M × N matrix. Second, the M × N matrix is provided to the FCA component which performs formal concept analysis to summarize the evolution of the entities along a temporal axis. Third, a summary merger policy is applied to each output of the FCA component to produce a set of abstract molecules. Each abstract molecule represents the temporal evolution of a single entity between the versions of the knowledge graphs taken as input. Fourth, said summary knowledge graph will go through a verbalization process that converts the summary graph to a readable and chronological text using a verbalization pipeline. Each summary molecule represents a single entity's temporal evolution over the knowledge graph versions taken as input.

3.3 Conversion of Knowledge Graphs to Groups of Equivalent RDF Molecules

The pipeline receives any number of KGs $\phi_1(D)$,...,$\phi_n(D)$ as input where $1, \ldots, n$ represent the different temporal versions of the same KG $\phi(D)$. First each graph is individually converted into sets of RDF molecules. Thus we obtain RDF molecule sets $S_1, \ldots, S_n$ which correspond to graphs $\phi_1(D), \ldots, \phi_n(D)$ respectively. The pipeline then identifies equivalent molecules within $S_1, \ldots, S_n$. As $\phi_1(D), \ldots, \phi_n(D)$ are different temporal versions of the same KG, it can be inferred that there exists equivalent molecules $M_1, \ldots, M_n$ such that $M_1 \in S_1, \ldots, M_n \in S_n$ and $M_1, \ldots, M_n$ all represent the same real-world entity. For the sake of simplicity it is assumed here that equivalent entities retain the same URI. Practically, semantic similarity measures as demonstrated in [2] can also be integrated with this pipeline to identify equivalent entities in cases the URI is different.

3.4 Applying Formal Concept Analysis to Obtain a Summary of Evolution

Formal concept analysis studies binary object-attribute tables to describe the relationship between *objects* and their *attributes*. Our approach first converts KGs to RDF molecules. Within a single KG, an RDF molecule can be considered as an *object* while its object or data properties can be considered as *attributes*. When RDF molecules are modeled in this way, we are able to apply the formal concept analysis algorithm to compute formal concepts.



1. Content Planning: Content Selection and Ordering. Ranking the triples of the summary with a chronological order	**2. Sentence Planning:** Sentence aggregation, Lexicalization, and Referring expression generation	**Final Text**
ALL, name, Cristiano Ronaldo . ALL, born in, Madeira . ... 2002, team, Sporting . 2008, team, Man. Unit . 2008, profession, Fashion Entrepreneur . 2008-2013-2020, language, English . ... 2013, team, Real Madrid . 2013, partner, Irina Shayk . 2013-2020, son, Cristiano Jr . ... 2020, team, Juve . 2020, partner, Georgina Rodriguez .	team → is player of occupation → is a language → speaks ... 4) In 2008, CR7 is player of Man United. 5) In 2008, CR7 is a Fashion Entrepreneur. 6) From 2008 to 2020, CR7 speaks English.	Cristiano Ronaldo was born in Madeira on the 5th of February of 1985. He is a Football Player. In 2002 Cristiano Ronald was playing for Sporting. In 2008 he was a player of Man United and started his career as a Fashion Entrepreneur, he as well learned English. In 2013, Cristiano Ronald was playing for Real Madrid and was dating Irina Shayk. He learned Spanish, started his modeling career, and got a son Cristiano Jr. Now he plays for Juve and is dating Georgina Rodriguez.
	3. Realization: Lexical rules for realization, Syntax / Grammar rules In 2008 (TE), CR7 (PN , he) is player of (VP, TENSE: PAST), and CR7 (Removed) is a (VP, TENSE: PAST) Fashion Entrepreneur, CR7 as well speaks (VP, TENSE: PAST) English.	

Fig. 3: **Verbalization Pipeline.**

In the previous step we obtained sets of molecules that correspond to different temporal versions of the same real-world entity, e.g., in our motivation example we refer to the life events of CR7. We apply V. Vychodil's algorithm [20] on each group of RDF molecules. The algorithm returns a set of formal concepts $< M, P >$ where M is a set of all the molecules that have all the properties contained in P. In our approach the output $< M, P >$ from formal concept analysis gives us a set of molecules that have the same properties throughout different KG versions. Following our motivating statement, we now can obtain the information that throughout the years 2002 and 2008 CR7 remained a football player and spoke Portuguese. Next, a summary fusion policy is applied to the output of the Formal Concept Analysis algorithm to obtain the temporal summary of all the different versions of the molecules.

3.5 Verbalizing the entity summary

In this step, we take as input the graph summary freshly produced, and generate a report of entity evolution. We define the task of verbalization as a function, and therefore it is possible to choose from the different approaches reviewed in Section 2. In this study, however, we explore a template-based approach, mainly because we have a controlled vocabulary providing us an exact phenomenon to verbalize: the evolution of an entity.

Following the Natural Language Generation (NLG) pipeline described by Reiter et al. [13], we divide the summary verbalization into three steps: 1) Content Planning, 2) Sentence Planning, and 3) Realization (see Figure 3). First, we plan the content by ordering the triples chronologically; the triples that cover all the years ranked at the top. Thus, we order triples from the oldest changes to more recent ones. As the second step, we start building sentences using lexicalization, and referring expressions. We assume that there are `rdfs:label` descriptive enough to plan sentences with relative readiness. For example to transform the relation `team` to "*is player of*". Additionally we define group of sentences that should be verbalize together to form a concise idea and message. Finally, in step three, lexical and grammar rules are used to produce

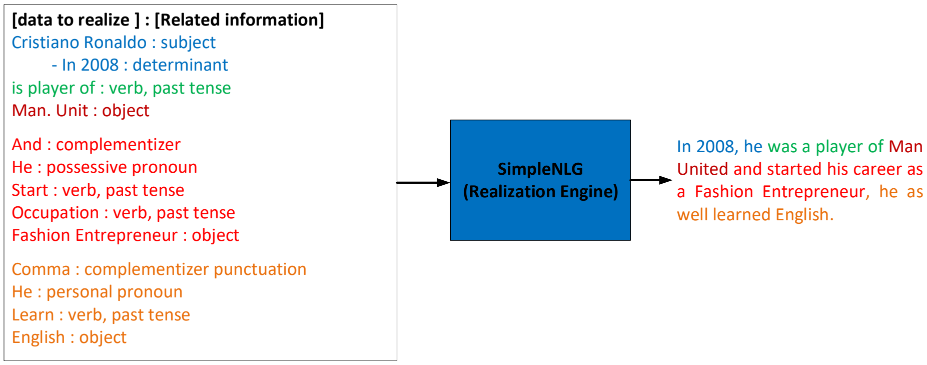

[data to realize] : [Related information]
Cristiano Ronaldo : subject
- In 2008 : determinant
is player of : verb, past tense
Man. Unit : object
And : complementizer
He : possessive pronoun
Start : verb, past tense
Occupation : verb, past tense
Fashion Entrepreneur : object
Comma : complementizer punctuation
He : personal pronoun
Learn : verb, past tense
English : object
SimpleNLG
(Realization Engine)
In 2008, he was a player of Man United and started his career as a Fashion Entrepreneur, he as well learned English.

Fig. 4: **Realization Engine.**

the final text. Thus, a Realization Engine is required at this step. We propose to use SimpleNLG [10]. SimpleNLG requires two elements: the data to realize, and related information. Figure 4 shows an example of how SimpleNLG works.

4 Conclusion and Future Lines of Work

In this article, we introduced our approach to verbalize the evolution of entities in KGs. Our approach leverages the concepts of RDF molecules, Formal Concept Analysis, Entity Summary, and a Verbalization Function. We explain the architecture and pipeline where only one parameter is needed, i.e., an entity filter. The report created can be useful in several domains e.g., manufacturing, healthcare, or financial domain.

Future lines of work. To date, several lines of research are ahead of us, including the performance and evaluation of the approach, and the use of FCA for QA systems. FCA has scalability limitations when applied to big knowledge graphs. Therefore a newly adapted version of FCA for Big Data scenarios needs to be employed. We need to define a fair evaluation framework for our approach, including different datasets and metrics, e.g., BLUE score. Finally, we believe the summaries produced by FCA may be useful for QA systems, for example, to answer evolutionary questions on knowledge graphs and answer questions about differences between entities in a Knowledge Graph.

Acknowledgements. We acknowledge the support of the EU projects Cleopatra (GA 812997) and TAILOR (GA 952215), the Federal Ministry for Economic Affairs and Energy (BMWi) project SPEAKER (FKZ 01MK20011A), the German Federal Ministry of Education and Research (BMBF) projects and excellence clusters ML2R (FKZ 01 15 18038 A/B/C), MLwin (01S18050 D/F), ScaDS.AI (01/S18026A), the Fraunhofer Zukunftsstiftung project JOSEPH, as well as the Marie Skłodowska-Curie fellowship programmes GA 801522 and GA 713567 at the ADAPT SFI Research Centre (grant 13/RC/2106).

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. (2015)
2. Collarana, D., Galkin, M., Ribón, I.T., Vidal, M., Lange, C., Auer, S.: MINTE: semantically integrating RDF graphs. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS. (2017) 22:1–22:11
3. Diefenbach, D., Giménez-García, J.M., Both, A., Singh, K., Maret, P.: Qanswer KG: designing a portable question answering system over RDF data. In: ESWC 2020 Proceedings. Volume 12123. (2020) 429–445
4. Diefenbach, D., Singh, K.D., Maret, P.: Wdaqua-core1: A question answering service for RDF knowledge bases. In: Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018, ACM (2018) 1087–1091
5. Endris, K.M., Faisal, S., Orlandi, F., Auer, S., Scerri, S.: Interest-Based RDF Update Propagation. In: The Semantic Web - ISWC 2015. Volume 9366. (2015) 513–529
6. Fernández, J.D., Umbrich, J., Polleres, A., Knuth, M.: Evaluating query and storage strategies for rdf archives. In: SEMANTiCS 2016. (2016)
7. Fernández, J.D., Debattista, J., Orlandi, F., Vidal, M.E.: Mepdaw chairs' welcome. In: Companion of The 2019 World Wide Web Conference, WWW. (2019) 13–17
8. Fernández, J.D., Llaves, A., Corcho, O.: Efficient rdf interchange (eri) format for rdf data streams. In: International Semantic Web Conference, Springer (2014) 244–259
9. Ferré, S., Cellier, P.: Graph-fca: An extension of formal concept analysis to knowledge graphs. Discret. Appl. Math. **273** (2020) 81–102
10. Gatt, A., Reiter, E.: Simplenlg: A realisation engine for practical applications. In: ENLG 2009 - Proceedings of the 12th European Workshop on Natural Language Generation, March 30-31, 2009, Athens, Greece. (2009) 90–93
11. González, L., Hogan, A.: Modelling dynamics in semantic web knowledge graphs with formal concept analysis. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, ACM (2018) 1175–1184
12. Nema, P., Shetty, S., Jain, P., Laha, A., Sankaranarayanan, K., Khapra, M.M.: Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In: NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018. (2018) 1539–1550
13. Reiter, E., Dale, R.: Building natural language generation systems. Cambridge Press (2000)
14. Rouane Hacene, A.M., Huchard, M., Napoli, A., Valtchev, P.: Relational Concept Analysis: Mining Concept Lattices From Multi-Relational Data. Annals of Mathematics and Artificial Intelligence **67**(1) (2013) 81–108
15. Roussakis, Y., Chrysakis, I., Stefanidis, K., Flouris, G., Stavrakas, Y.: A flexible framework for understanding the dynamics of evolving RDF datasets. In: ISWC 2015 - International Semantic Web Conference. Volume 9366. (2015) 495–512
16. Singh, A., Brennan, R., O'Sullivan, D.: DELTA-LD: A Change Detection Approach for Linked Datasets. In: 4th MEPDaW Workshop at ESWC. (2018)
17. Tasnim, M., Collarana, D., Graux, D., Galkin, M., Vidal, M.: COMET: A contextualized molecule-based matching technique. In: International Conference on Database and Expert Systems Applications, Springer (2019) 175–185
18. Tasnim, M., Collarana, D., Graux, D., Orlandi, F., Vidal, M.: Summarizing entity temporal evolution in knowledge graphs. In: Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, ACM (2019) 961–965
19. Umbrich, J., Villazón-Terrazas, B., Hausenblas, M.: Dataset Dynamics Compendium: A Comparative Study. In: 1st Workshop on Consuming Linked Data (COLD2010). (2010)
20. Vychodil, V.: A new algorithm for computing formal concepts. na (2008)