
EventNarrative: A Large-scale Event-centric Dataset for Knowledge Graph-to-Text Generation (Supplementary Material)

Anthony Colas^{*}, Ali Sadeghian^{*}, Yue Wang, Daisy Zhe Wang
Department of Computer Science, University of Florida
{acolas1, asadeghian, yue.wang1, daisyw}@ufl.edu

The supplementary material consists of two parts: the dataset documentation and intended uses which includes collection, distribution, maintenance, and licensing details (Section A) and an example illustrating each step in the dataset creation approach (Section B).

A Dataset documentation and Intended Uses

We follow the dataset documentation framework provided by datasheets for datasets [1].

Our dataset can be found at: <https://www.kaggle.com/acolas1/eventnarration>.

A.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on event-centric knowledge graph-to-text: given an event-centric knowledge graph (KG), narrate the graph by using natural language sentences. It was also created as the first large-scale parallel knowledge graph-to-text dataset with a rich ontology. To the best of our knowledge, no such dataset existed previously.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Anthony Colas, Ali Sadeghian, and Daisy Wang from the Data Science Research Lab at the University of Florida.

Who funded the creation of the dataset?

Anthony Colas: Graduate School Preeminence Award (GSPA) at the University of Florida, as well as the McKnight Doctoral Fellowship. Daisy Wang: NSF under IIS Award #1526753 and DARPA under Award #FA8750-18-2-0014(AIDA/GAIA)

Any other comments?

N/A

A.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances represent real-world events as graph and text (i.e., event-centric KG-narrative pairs). Event KGs are extracted from EventKG [2] and Wikidata [3] and are linked to their Wikipedia text.

An example is shown in Section B.

How many instances are there in total (of each type, if appropriate)?

The dataset 224,428 instances (KG-narrative pairs) in total.

Does the dataset contain all possible instances or is it a sample(not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances which were found to be linked between the event KGs and Wikipedia narrative, containing the criteria described in our Dataset Creation approach. Therefore, it is a sample of the events found in EventKG. No tests were run to determine representativeness.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The data consists of KGs represented as (subject, predicate, object) triples, the narrative represented as natural language text, and the entities found in the text as a dictionary.

Is there a label or target associated with each instance? If so, please provide a description

Each KG has a target narrative in raw text format. Depending on the task, the text can be used as labels for the graphs and vice-versa.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, but individual instances contain event types which may be shared, e.g., sports season, battle, etc. Other relationships between individual instances may appear within their common KG components.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The dataset is released with its specified 80/10/10 train/development/test split such that none of the events in the training split are in the development and test split and vice versa. The dataset also has a “small development” split made up of 1,000 pairs for use by models which are computationally expensive in the generative step, which can be found on our website.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Not that the authors are aware of.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals’ non-public communications)? If so, please provide a description.

The dataset does not contain any confidential data.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description

Any other comments?

A.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was all observable as raw text and collected from EventKG (<http://eventkg.13s.uni-hannover.de/>), Wikidata (https://www.wikidata.org/wiki/Wikidata:Main_Page), and Wikipedia (<https://www.wikipedia.org/>). 500 randomly sampled instances were manually validated as described in Section 4.3 (Qualitative Analysis).

What mechanisms or procedures were used to collect the data(e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

All data were automatically extracted from EventKG, Wikidata, and Wikipedia and further processed via the mechanisms described in Section 3 (Dataset Creation). Validation was done via qualitative analysis on a random sample of 500 instances as described in Section 4.3 of our paper.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset collection process began with all events from EventKG which had a Wikidata and Wikipedia page. It was further reduced given the procedures and rules described in Section 2 (Dataset Creation).

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Student annotators voluntarily verified the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time-frame in which the data associated with the instances was created.

The final Wikidata related data was retrieved from the Wikidata SPARQL endpoint on May 2021. We hosted a local copy of Wikipedia from `wikipedia_en_all_nopic_2021-01.zim`, which can be found here ¹. For EventKG data, EventKG 1.0 was used.

¹<https://www.mirrorservice.org/sites/download.kiwix.org/zim/wikipedia/>

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical review processes were conducted.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

All data was collected from EventKG, Wikidata, and Wikipedia as mentioned above.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

All data was publicly available.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A (see previous question).

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A.

Any other comments?

A.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Instances found in EventKG which did not have both a Wikidata resource page and Wikipedia page ID were discarded. From these, those for which we could not find any paragraph-related text on Wikipedia were discarded. We then discarded those instances for which no entities were found in the text. Further preprocessing was done via entity matching and graph generation as described in Section 2 of the corresponding paper. In order to ensure that the event narratives were not too lengthy, all events with more than 500 tokens in their text were discarded.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data

Yes, the “raw” data was saved. Unprocessed files can be found here. Data at specific stages of the dataset creation process will be made available upon request.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

We have posted all code used to preprocess the data on GitHub:
<https://github.com/acolas1/EventNarrative>.

[Any other comments?](#)

A.5 Uses

[Has the dataset been used for any tasks already?](#)

At the time of this submission, the dataset has only been used for the knowledge graph-to-text task.

[Is there a repository that links to any or all papers or systems that use the dataset?](#)

We will maintain a repository at:

<https://github.com/acolas1/EventNarrative>.

[What \(other\) tasks could the dataset be used for?](#)

The dataset could be used for any task related to event-centric KGs and their corresponding text. For example, our dataset could be used to train models involving event-centric entity linking and relation extraction.

[Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups \(e.g., stereotyping, quality of service issues\) or other undesirable harms \(e.g., financial harms, legal risks\) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?](#)

No, there is minimal risk from any harm as all the data sources are publicly and freely available. Wikidata and Wikipedia are made available under the *Creative Commons CC0 License* and EventKG under the *CC BY 4.0 license*.

[Are there tasks for which the dataset should not be used? If so, please provide a description.](#)

The dataset is event-centric and should be used on event-related tasks. We ask that our dataset not be used to create fabricated news or spread misinformation.

[Any other comments?](#)

A.6 Distribution

[Will the dataset be distributed to third parties outside of the entity \(e.g., company, institution, organization\) on behalf of which the dataset was created? If so, please provide a description.](#)

Yes, the dataset is publicly available on the internet.

[How will the dataset be distributed \(e.g., tarball on website, API, GitHub\)? Does the dataset have a digital object identifier \(DOI\)?](#)

The dataset will be initially distributed on Kaggle, via <https://www.kaggle.com/acolas1/eventnarration/>, afterwards it will be distributed on the University of Florida's Data Science Research's Lab website. For the code, please see <https://github.com/acolas1/EventNarrative>. The dataset will have a DOI once it becomes public.

[When will the dataset be distributed?](#)

The dataset will be publically distributed after the NeurIPS 2021 camera-ready deadline.

[Will the dataset be distributed under a copyright or other intellectual property \(IP\) license, and/or under applicable terms of use \(ToU\)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU,](#)

as well as any fees associated with these restrictions.

The dataset is distributed under the *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication* license. We request that any use of the EventNarrative dataset cite the corresponding paper.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown.

Any other comments?

A.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset will be supported, hosted, and maintained by the Data Science Research Lab at the University of Florida for up to 10 years.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? All questions and comments about the dataset can be sent to Anthony Colas: acolas1@ufl.edu. Other contacts can be found at: dsr.cise.ufl.edu/people/.

Is there an erratum? If so, please provide a link or other access point.

All changes to the dataset will be announced on the website from which we will temporarily host the dataset ², the University of Florida Data Science Lab's website ³, as well as the dataset's corresponding GitHub page⁴.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes. Any updates to the dataset will be communicated via GitHub and the University of Florida's Data Science Research Lab's website and the dataset's corresponding GitHub page. Updates are planned to be done by the authors semi-annually. Updates include adding new events as well as adding new properties to existing events.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

All versions will be continued to be hosted and maintained with information posted on the Data Science Research Lab at the University of Florida's website and dataset's GitHub page. We will support the latest version, while providing limited support to earlier versions of the dataset. The initial version is hosted on <https://www.kaggle.com/acolas1/eventnarration>.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism

² <https://www.kaggle.com/acolas1/eventnarration>

³ <https://dsr.cise.ufl.edu/>

⁴ <https://github.com/acolas1/EventNarrative>

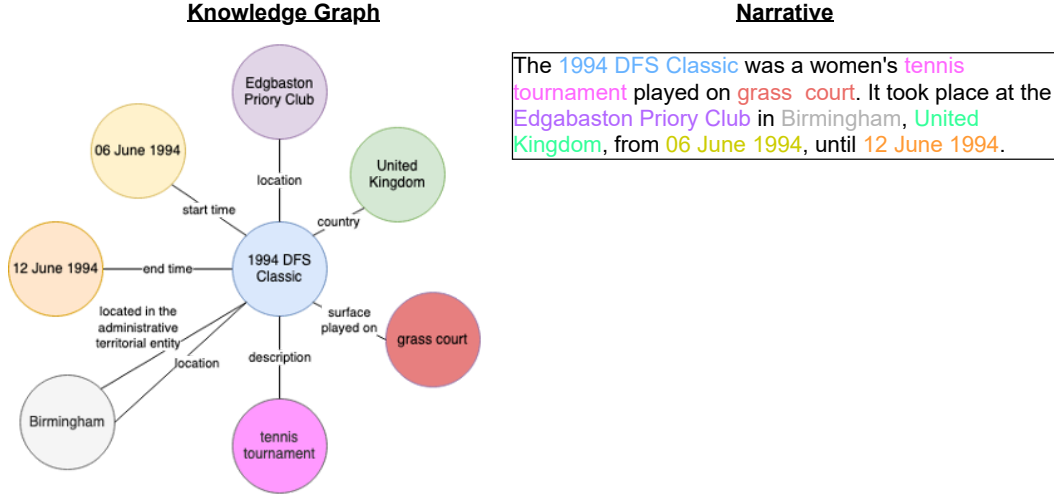


Figure 1: An example KG-Narrative pair for the 1994 DFS Classic event. Corresponding entities found in the KG and narrative are color coded.

for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, others may do so and should contact the original authors about extensions to the datasets for which they wish to be hosted on the website. These contributions will be validated/verified via thorough communication with those who wish to extend the dataset. These contributions will be posted as an announcement via the website which hosts the dataset.

Any other comments?

A.8 Responsibility

The authors bear all responsibility in case of violation of rights, etc., and confirm that the dataset is released under the *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication* license.

B Example Workflow

B.1 Example

We now illustrate the dataset creation approach used for EventNarrative using illustrative examples. Throughout the forthcoming steps, we do so by showcasing the *1994 DFS Classic* event, a tennis tournament. We illustrate separate and isolated examples for stages which may involve more intricate processes, which do not apply to the *1994 DFS Classic*, i.e., the step within *Sources* which involves joining multiple knowledge graphs (KGs). See Figure 1 below for the final processed event KG-narrative pair corresponding to the *1994 DFS Classic*.

B.2 Sources

As detailed in the Data Creation section of our work (Section 2), we source our original KGs from EventKG and augment the graphs with additional data found in Wikidata. For the *1994 DFS Classic*, both graphs are mostly identical as shown in Figure 2. However, in this step we normalize relations such as the *end time*, *start time*, *location*, *next event*, *previous event* while removing those nodes which only describe temporal metadata and ontological identifiers. For example, in EventKG many events have relation *hasBeginTimeStamp* and *hasEndTimeStamp* as *YEAR-01-01* and *YEAR-12-31*,

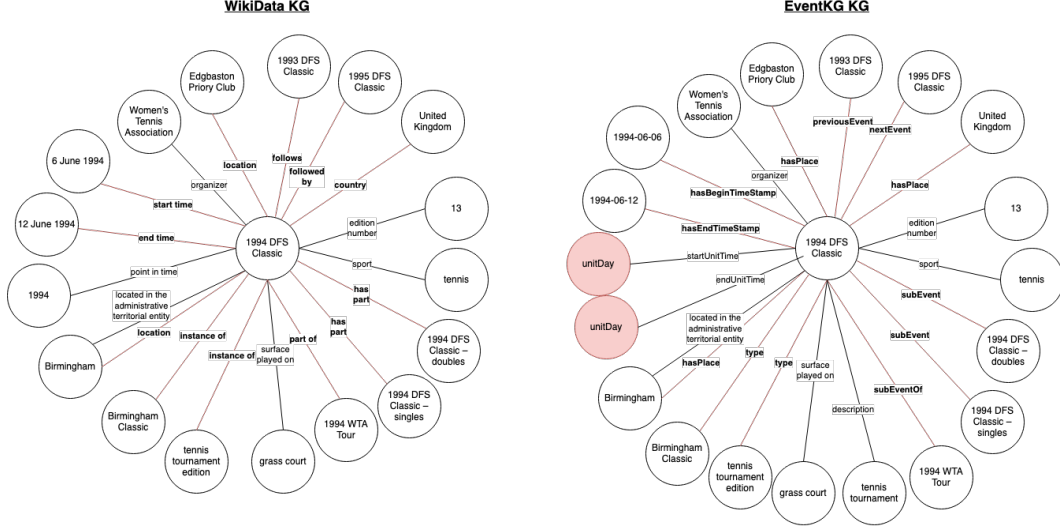


Figure 2: Left: The Wikidata graph corresponding to the 1994 DFS Classic. Right: The EventKG graph corresponding to the 1994 DFS Classic. Relations shared by both Wikidata and EventKG, which contain different labels, are highlighted in red. Nodes containing temporal metadata are highlighted in red. Note, we omit some information, such as metadata, for brevity.

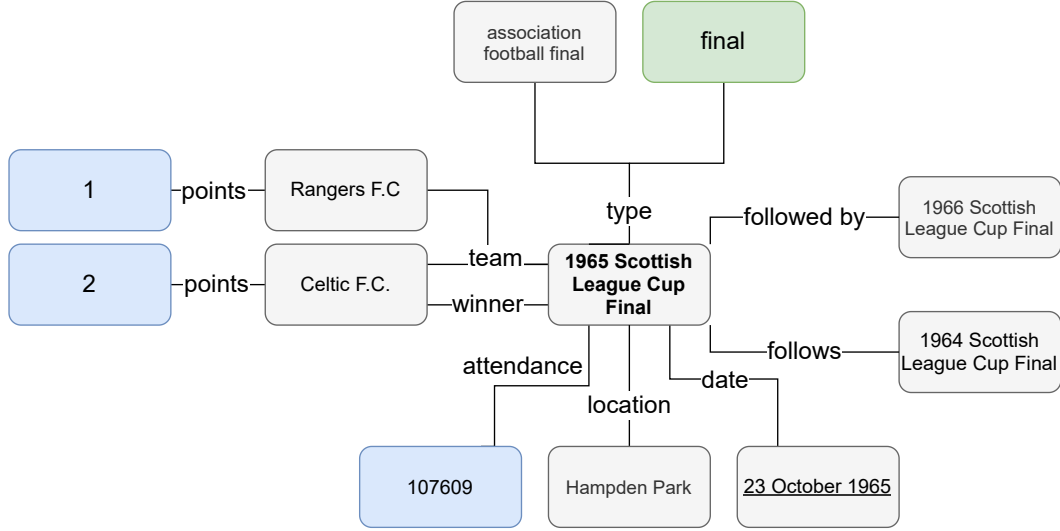


Figure 3: An example sub-graph for the 1965 Scottish League Cup Final, depicting entities/nodes found only in Wikidata (in blue) and found only in EventKG (in green). Note, to emphasize that this graph is different from the example above, we depict the nodes as rectangles.

respectively, specifying that the event is a yearly event, and associated with the *point in time* relation found in Wikidata.

See Figure 3, depicting the *1965 Scottish League Cup Final*, for a KG containing entities found exclusively in either Wikidata or EventKG.

B.3 Full Text Merge

To connect a joined KG with its corresponding narrative, we utilize the *wikipediaLabel* from EventKG and retrieve the event's corresponding Wikipedia article. We do so by hosting a local copy of Wikipedia from *wikipedia_en_all_nopic_2021-01.zim* which can be found here:

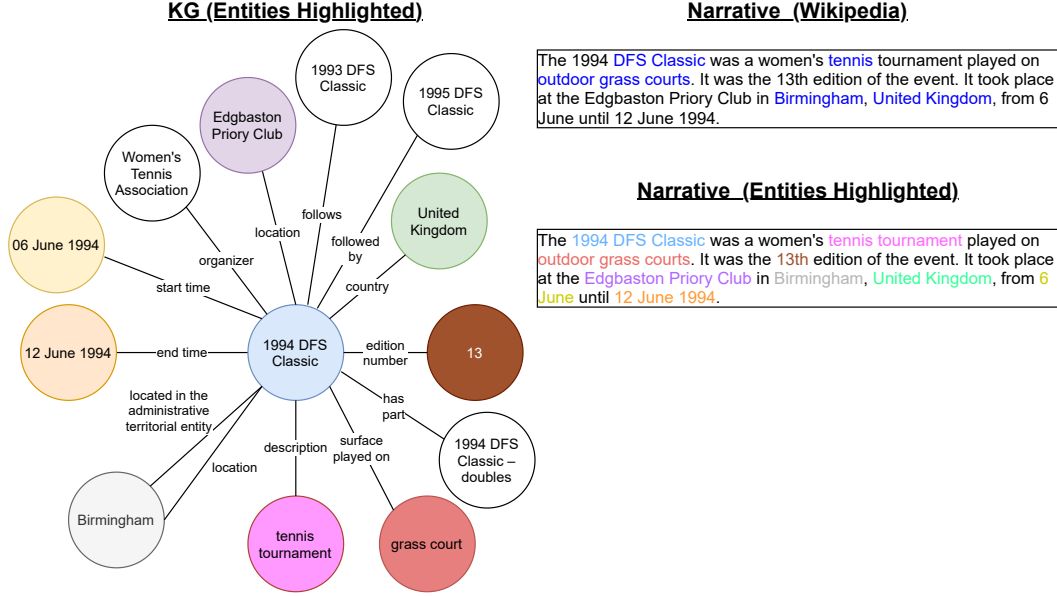


Figure 4: **Left:** The 1994 DFS Classic KG with its corresponding color coded entities in the narrative. **Right (top):** The Wikipedia article text for the 1994 DFS Classic. Note the highlighted Wikipedia hyperlinks. The QID for these pages are retrieved in the QID matching step. **Right (bottom):** The Wikipedia article text for the 1994 DFS Classic, with the corresponding matched entities in the text highlighted with their respective color.

<https://www.mirrorservice.org/sites/download.kiwix.org/zim/wikipedia/> . For those articles not found, we try calling the online Wikipedia API.

B.4 Entity Matching

Figure 4 illustrates an example for Entity Matching on the 1994 DFS Classic event. Namely, the hyperlink/QID, exact, and date matching steps. In the QID matching step, each hyperlink's QID is extracted and matched with the retrieved Wikidata item's QID. From the example, one can also see the longest match process take effect, where a match for *tennis tournament* instead of *tournament* is found.

B.5 Narrative Graph Generation

Finally, in the Narrative Graph Generation step we reduce the event's KG and narrative recursively, found in Figure 4 until each sentence in the narrative contains at least two entities. We summarize our Narrative Graph Generation process in the following manner:

1. Remove any sentences with less than two entity matches.
2. If a sentence containing an entity match is removed, and the entity is no longer contained in the text, we remove the entity from the KG.
3. If any disconnected graphs are obtained after removing the entity, we discard the disconnected components from the KG and remove their corresponding entity from the narrative.
4. Repeat this process recursively.
5. Stop if there is no change in the KG or text to produce the event KG-narrative pair.

A simple example can be seen in Figure 5, where the sentence "It was the 13th edition of the event." and its parallel KG component are removed, because it contains less than two matched entities.

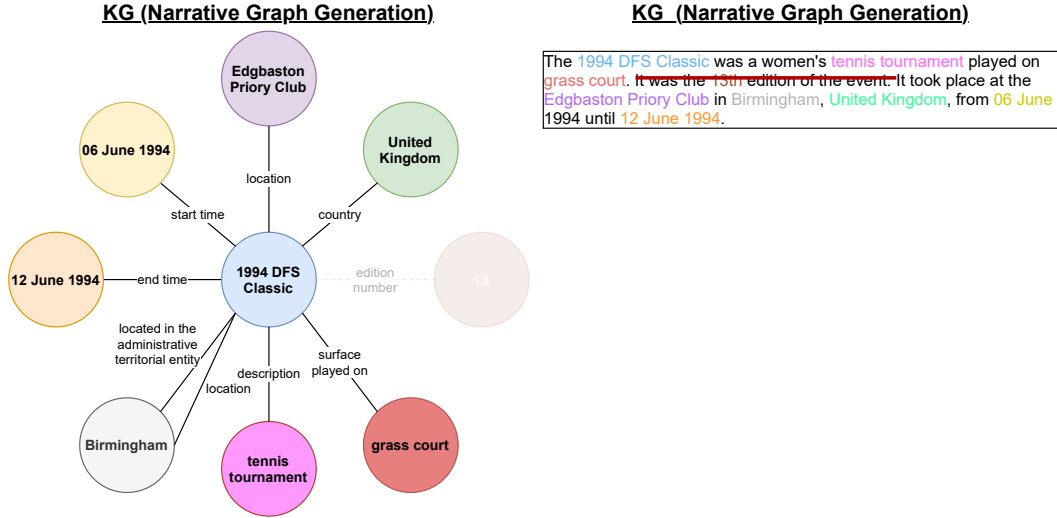


Figure 5: An illustration of the Narrative Graph Generation step, where the sentence and KG elements corresponding to the node labeled “13” are removed.

References

- [1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, III Hal Daumé, and Kate Crawford. Datasheets for datasets. 2018.
- [2] Simon Gottschalk and Elena Demidova. Eventkg: A multilingual event-centric temporal knowledge graph. In *European Semantic Web Conference*, pages 272–287. Springer, 2018.
- [3] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.