WDV: A Broad Data Verbalisation Dataset Built from Wikidata

Gabriel Amaral(✉), Odinaldo Rodrigues, and Elena Simperl

King's College London, London WC2R 2LS, UK
{gabriel.amaral,odinaldo.rodrigues,elena.simperl}@kcl.ac.uk

Abstract. Data verbalisation is a task of great importance in the current field of natural language processing, as there is a clear benefit in the transformation of our abundant structured and semi-structured data into human-readable formats. Verbalising Knowledge Graph (KG) data focuses on converting interconnected triple-based claims, formed of subject, predicate, and object, into text. Although KG verbalisation datasets exist for some KGs, there are still limitations in their applicability to many scenarios. This is especially true for Wikidata, where available datasets either loosely couple claim sets with textual information or heavily focus on predicates around biographies, cities, and countries. To address these gaps, we propose WDV, a large KG claim verbalisation dataset built from Wikidata, with a tight coupling between triples and text, covering a wide variety of entities and predicates. We also evaluate the quality of our verbalisations through a reusable workflow for measuring human-centred fluency and adequacy scores. Our data (https://doi.org/10.6084/m9.figshare.17159045.v1) and code (https://github.com/gabrielmaia7/WDV) are openly available in the hopes of furthering research towards KG verbalisation.

Keywords: Crowdsourcing · Knowledge graphs · Data verbalisation

1 Introduction

Data verbalisation, a facet of Natural Language Generation (NLG), is a task that has great importance in the current field of natural language processing [10,14,15,32,35,44], as there is great benefit in the transformation of our abundant structured and semi-structured data into human-readable formats. It is important in its own right, as well as as a step toward larger tasks such as open-domain question-answering [23] and automated fact checking [40,41]. One large source of semi-structured data that would benefit greatly from verbalisation is collaborative Knowledge Graphs (KG) like DBpedia[1] and Wikidata.[2]

The verbalisation of KGs data consists of converting sets of claims into natural language text. Each claim consists of a triple, formed of subject, predicate,

[1] https://www.dbpedia.org/.
[2] https://www.wikidata.org.

U. Sattler et al. (Eds.): ISWC 2022, LNCS 13489, pp. 556–574, 2022.
https://doi.org/10.1007/978-3-031-19433-7_32

and object, and each claim set shares subjects and objects; the verbalisation then has to deal with expressing and linking these pieces of information. Although KG verbalisation datasets, mapping claim sets to text, exist for some popular KGs [2,6,11], they are not without their limitations.

Wikidata, the web's largest collaborative KG, has very few such datasets [6, 38], and existing ones rely on distant supervision to prioritise the sheer number of couplings in exchange for coupling tightness. They also disproportionately represent specific entity types from Wikidata (e.g. people and locations) when Wikidata covers a much wider variety of information. Finally, data verbalisation performance is mainly measured via algorithms, such as BLEU [27], which have been the target of many criticisms when applied to NLG [25,31,33].

We propose WDV, a large KG verbalisation dataset with 7.6k Wikidata entries. WDV addresses limitations in coverage disproportion, verbalisation coupling, and algorithmically measured quality, respectively, given that:

1. WDV is built from a much wider variety of entity types and predicates than similar datasets, and is intended as a benchmarking dataset for data verbalisation models applied on Wikidata;
2. WDV supports a tight coupling between single claims and text directly associating a triple-based claim and a natural language sentence;
3. 1.4k entries of WDV have been annotated by a collective of humans, allowing for the evaluation and future improvement of our verbalisations, as well as establishing a non-algorithmic baseline for other verbalisation models.

Additionally, we create a reproducible crowdsourcing workflow for capturing human evaluations of fluency and adequacy in graph-to-text NLG. All used code and gathered data are available in this paper's GitHub repository.

The remainder of the paper is structured as follows. Section 2 positions our dataset in regards to existing datasets. Section 3 presents our dataset construction, including quality annotations. Section 4 describes our verbalisation model. Finally, Sects. 5 and 6 reinforce and summarise our contributions.

2 Background and Related Work

Verbalising KGs consists of generating grammatically correct natural language based on structured and semi-structured data from a KG, maintaining the original meaning. This data is encoded in triples (claims), consisting of a subject, a predicate, and an object; all three components model aspects of knowledge, such as entities, classes, attributes, and relationships. Examples of popular KGs are DBpedia, Wikidata, Yago,[3] and Freebase.[4] Their verbalisation is an important task on its own, but is also a key step in downstream tasks [23,36,40,41].

Datasets that align KG claims to text are vital for creating and evaluating KG verbalisation approaches. While several have been created, they are not

[3] https://github.com/yago-naga/yago3.
[4] https://developers.google.com/freebase.

without their limitations. The *NYT-FB* [24,42] dataset aligns text from the New York Times with triples from Freebase through named entity linking and keyword matching against Freebase labels. This leads to a disproportional coverage of news-worthy entities and topics, such as geography and politics, and from a specific period in time, limiting its usefulness in broader scenarios. The same narrow scope is seen in the *TACRED* dataset [43], which covers only 41 relationships about people and organisations, such as age, spouse, shareholders, etc., as its data does not stem from any specific KG, but rather annotated newswire and web text from the TAC KBP corpora [5]. Also, its texts often contain much more information than their aligned triples, making it a resource not fully suited for NLG. The *FB15K-237* dataset [34] aligns Freebase triples to synsets instead of text, making it unusable for NLG without text grounding. Additionally, both NYT-FB and FB15K-237 rely on Freebase, which was discontinued and its data moved to Wikidata [28], compromising these datasets' usability and upkeep.

More recent datasets attempt to remedy some of these limitations. Pavlos et al. [37,38] propose two large corpora that align Wikidata and DBpedia claims to Wikipedia text. However, they focus on verbalisations of multiple claims at a time, which limits its usefulness for important tasks e.g. automated fact-checking in favour of others e.g. summarisation. Even more critically, they are based on distant supervision techniques, providing a loose alignment between sets of triples and text; triple sets consist of numerous claims that are very likely - but not guaranteed - to be expressed in the text, and the text contains information that is not assured to exist in the claims. The same is true for *T-REx* [6], which aligns Wikidata claims to Wikipedia abstracts, making it unreliable for NLG from KG claims while perfectly preserving their sense. Our dataset bridges this gap by focusing on a tight alignment between Wikidata claims and text.

The coverage issue seen in NYT-FB and TACRED is also present, although less so, in T-REx. It covers many unique predicates, yet they are disproportionately represented: the top 7.7% of its unique predicates represent 90% of its unique triples, and these mostly express information on people and places, with the `country` predicate alone representing over 11% of triples. The *WebNLG* [11] dataset remedies this by defining a list of very broad DBpedia classes and then collecting separate and balanced sets of claims from entities in each class. However, WebNLG also focuses on sets of multiple claims at a time.

We follow WebNLG's approach to resolving predicate and theme bias. However, we build WDV out of Wikidata instead, expanding the entity classes defined by WebNLG, as Wikidata lacks verbalisation datasets that cover its wide range of predicates and themes. To provide a better view of how WDV compares to other datasets mentioned in this Section, refer to Table 1.

3 WDV: An Annotated Wikidata Verbalisation Dataset

This section describes the construction of the WDV dataset, including crowdsourced annotations, as well as details of its structure. Figure 1 illustrates the entire process with numbered steps, which we cover in this Section. In a nutshell, it consists of first defining 20 large pools of filtered Wikidata claims, each

Table 1. Comparison between WDV and other KG verbalisation datasets. 'Entity Classes' shows in how many distinct themes the claims might be organised by, if at all. 'Text Alignment' refers to whether all text corresponds to aligned triples (Tight) or not (Distant). *Avail.* stands for Availability.

	Source graph	Aligned documents	Unique predicates	Unique triples	Entity classes	Text alignment	Avail
NYT-FB	Freebase	1.8M	258	39K	n.a	Distant	Partial
TACRED	n.a	106K	41	21K	n.a	Distant	Closed
FB15K-237	Freebase	2.7M	237	2.7M	n.a	Tight	Public
T-REx	Wikidata	6.2M	642	11M	n.a	Distant	Public
WebNLG	DBpedia	39K	412	3.2K	16	Tight	Public
WDV	Wikidata	7.6K	439	7.6K	20	Tight	Public

corresponding to a Wikidata class (steps 1–4). Then, we obtain a sample of claims from each pool such that predicates are represented as equally as possible (step 5). Lastly, we obtain aligned verbalisations and human annotations (steps 6 and 7). Throughout this entire construction process, data was extracted from a Wikipedia JSON dump from August 2021. The JSON format was used since the later stages of the pipeline i.e. crowdsourcing and verbalisation either require or greatly benefit from that input format. We also release WDV in this format as it targets ML practitioners and developers, who are very familiar with it.

To improve comprehensibility, transparency, and repeatability, we follow two recently proposed sets of guidelines. The first, by Gebru et al. [13], pertains to the effective documentation of machine learning datasets, supporting the transparency and reproducibility of their creation process. The second, by Ramirez et al. [30], pertains to the detailing of crowdsourcing experiments to guarantee clarity and repeatability. It ensures the impact of task design, data processing, and other factors on our conclusions, as well as their validity, can be assessed.

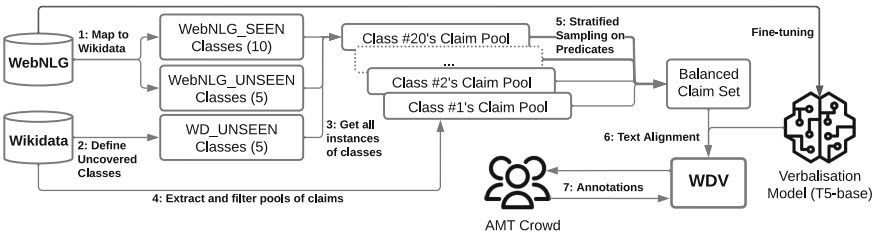


Fig. 1. Overview of WDV's construction workflow, starting with WebNLG's structure and Wikidata's contents, finishing with WDV and crowd annotations.

3.1 Balanced Claim Set Collection

WDV adapts and expands on WebNLG's partition and class structure to fit Wikidata. Firstly, this ensures a balanced representation of Wikidata entities

and predicates of various natures. Secondly, our data verbalisation model, used later in the workflow, is fine-tuned with WebNLG; keeping the same class composition thus reduces the chances of low-quality verbalisations. WebNLG has two partitions: SEEN, with 15 classes, and UNSEEN, with five, as seen in Table 2.

We start by mapping WebNLG's 15 DBpedia classes to their Wikidata equivalents (**step 1**). Some of Wikidata's most populous classes are not at all covered by these 15. Thus, from these uncovered classes, we select the five largest to compose an additional partition WD_UNSEEN (**step 2**); we do not consider ontological or scientifically complex classes (e.g. proteins). Next, we extract from Wikidata all entities that can be considered as instances or examples of these 20 classes or their subclasses (**step 3**), composing 20 large groups of entities.

From each class' extracted group of entities, we retrieve all claims that we deem suitable for verbalisation, based on the following exclusion criteria (**step 4**): we exclude deprecated claims, as they might contain incorrect or invalid values; claims with objects of datatypes that are hard to represent in natural language are excluded e.g. external database identifiers, URLs, images, mathematical formulas, etc.; we exclude claims that serve taxonomic or ontological purposes e.g. `subclass of (P31)`, `Topic's main category (P910)`, `See also (P1659)` etc.; and finally, claims whose objects are the special values *no value* or *some value*. The claims remaining after these exclusions compose 20 distinct pools of claims, or *themes*, from which we will next sample a set of claims.

These themes have very unbalanced distributions of claims over predicates e.g. over 50% of the claims in the *Airport* and *Mountain* themes have the `patronage (P3872)` and `country (P17)` predicates, respectively. A simple random sample would build a dataset that ignores the vast majority of Wikidata predicates. Hence, we opt for a stratified sampling of claims (**step 5**).

For each theme `t`, we determine the representative sample size N_t needed, considering its total number of claims, a 95% confidence interval, and a 5% margin of error. We start the sampling process by grouping each theme's claims by predicate, discarding very rare predicates (0.3% to 1.7% of total claims in a theme), and defining each theme's remaining M_t predicate groups as a stratum. For each theme `t`, we attempt to sample an equal amount of claims (N_t/M_t) from each stratum. If a stratum in theme `t` has less than N_t/M_t claims, we select all its claims and compensate by oversampling other strata in `t`, so that total sample size is still N_t. We keep track of all sampling weights in order to adjust any estimated statistic to account for the stratification. The resulting balanced claim set consists of statistically representative sets of claims from all 20 themes (7.6k claims in total), where predicates are as equally present as possible.

3.2 Text Alignment

WDV tightly aligns each claim to a natural language text i.e. each claim corresponds exactly to one sentence (and vice-versa), such that both hold the same meaning and the sentence is grammatically well-written. This is so that NLG is directly supported (as explored in Sect. 2) and also because WDV is the first step towards future research into automating fact checking for Wikidata.

To achieve this alignment (**step 6**), we first collect subject, predicate, and object labels (preferably in English) for each claim in the balanced claim set. We also collect aliases and descriptions, which play a part later in crowdsourcing. The collection is done by querying Wikidata's SPARQL engine.[5] In cases such as timestamps and measurements with units, label templates are used.

For each claim, its three labels are given to a verbalisation model, which outputs an English sentence that attempts to communicate the same information. The model itself, including its training and validation, is detailed in Sect. 4. This results in 7.6k claim-verbalisation pairings.

These claim-verbalisation pairings, alongside ontological attributes and the aggregated crowdsourced annotations (see Sect. 3.3), constitute WDV. Its detailed structure, an exemplary record, and some descriptive statistics are given in Sect. 3.4. Section 3.5 explores insights obtained from crowd annotations.

3.3 Crowdsourced Annotations

To measure how much of the claims' meanings are kept (i.e. adequacy) by the verbalisations and how much they resemble text written by humans (i.e. fluency), as well as to support the dataset's refining and correction, we crowdsource human annotations (**step 7**). These annotations are collected for a portion of WDV (20% of total claims) due to budget constraints, randomly selected among those claims having all labels in English, while keeping a proportional representation of each theme. Claim components not labelled in English are a minority that would represent a hurdle for crowd workers [22] and bias results.

Experimental Design. Before crowdsourcing, the WDV data goes through two pre-processing steps: *golden data generation* and *task composition*. Golden data is a small data subset that is manually annotated and used as a reference to discern between good and bad workers. We calculate how much golden data is necessary by minimizing, based on available data from similar studies [1], the probability of a regular worker finding a repeated set of golden data in two different tasks, which plateaus near 100% with 90 golden data annotations.

We take 45 random records from the sampled WDV data and set them aside as golden data for both fluency and adequacy tasks. We manually generate another 90 uniquely identified pairs to represent poor model performance: 45 for the fluency task by writing different levels of gibberish, and 45 for adequacy by randomly shuffling their verbalisations. We annotate golden data by defining, for each pair, what would represent reasonable scores for fluency and adequacy.

Task composition consists of: first, grouping the sampled WDV data plus the golden data such that each group (a *task set*) has two random golden data pairs and four random non-annotated pairs; then, attributing to each task a unique identifier; and lastly, sending the task set to the crowd embedded in an HTML script to be solved by at least five different workers.

[5] https://query.wikidata.org/.

Pilots were run in August 2021, and main tasks were run between September and October of the same year. Pilots helped us measure median time spent by workers to define fair payment, and collect general feedback to adjust task design. We calculated pay based on double the US's minimum hourly wage of USD7.25, in order to properly account for good workers that need more time than the median. We paid USD0.50 per fluency task and USD1.00 per adequacy task. Workers rated our tasks as having fair pay on TurkerView.[6] Before starting the task, workers are made aware of the pay and conditions and are told that continuing with the task means consenting to both.

Crowd. Crowd workers were selected from Amazon Mechanical Turk (AMT), the demographics of which have been explored in several papers [3,4,18]. We limited the tasks only to workers that had a good grasp of English by including an English grammar screening quiz before each task. Secondly, we only allowed workers that had done over 1000 tasks with over 80% acceptance rate to work on our tasks. We analysed contributions from the pilot, identifying workers that exhibited malicious behaviour and banning them from the main tasks.

Tasks. Task sets are sent to be annotated embedded in HTML pages. There is one for *fluency* and one for *adequacy* annotation tasks. Before starting either task type, workers are shown a description of that task, rules, and instructions they should follow. They also see many examples of acceptable answers with explanations. Workers can access this information at all times during the task.

In the fluency task, workers are shown only the verbalisation and are asked to rate its fluency with a score from 0 to 5, 0 being the worst and 5 being the best. In the adequacy task, workers are shown both the verbalisation and the claim, as well as labels, aliases, and descriptions, and are asked whether they convey the same information. They can reply *Yes* (giving it a score of 0), *No* (score of 1), and *Not Sure* (score of 2). Answering *No* and *Not Sure* prompts a question as to the reason; workers can blame the verbalisation, each component in the triple, a combination, or select *Other* and give a new justification. These tasks were released on AMT after receiving ethical approval.

Quality Control. Multiple quality control techniques were applied. The small randomized grammar quiz at the start of the task serves as an attention check, discouraging spammers. Our gold data is used to measure worker quality during the task, alongside other checks such as time spent per pair and whether all questions were answered. Failing these checks alerts the user and asks them to reevaluate their annotations. Failing three times closes the task without submission. Workers are told these details before engaging with the task.

Task Code and Raw Data. All the code and data for our crowdsourcing are in this paper's GitHub repository, including detailed descriptions of each task's

[6] https://turkerview.com/.

execution and the exact HTML code sent to each anonymous worker alongside instructions, agreement terms, and examples. It also includes all retrieved data before it was processed and aggregated back into WDV.

3.4 WDV Composition

WDV consists of a large partially annotated dataset of over 7.6k entries that align a broad collection of Wikidata claims with their respective verbalisations. An example of an annotated record can be seen in Fig. 2. The attributes seen there consist of: attributes describing the claim, such as its Wikidata ID (*claim_id*) and its *rank* (normal, deprecated or preferred); attributes from the claim's components (subject, predicate, and object), including their Wikidata IDs (e.g. *subject_id*), labels (e.g. *subject_label*), descriptions (e.g. *subject_desc*), and aliases (e.g. *subject_alias*); a JSON representation of the *object* alongside its type (*object_datatype)* as defined by Wikidata; attributes from the claim's theme such as its root class' Wikidata ID (theme_root_class_id) and label (theme_label); the aligned *verbalisation*, before and after replacement of tokens unknown to the model (*verbalisation_unk_replaced*); the *sampling weight* from the stratified sampling process; and the crowdsourced *annotations* and their aggregations, for those entries (∼1.4k) that are annotated.

Our schema is different from the Wikipedia dumps' JSON schema. Firstly, the latter is entity-centered: each entry is an entity and claims are components hierarchically encoded as elements. As WDV is centered on claim-verbalisation alignments, we flatten this structure. Secondly, information on the claims' components is spread over their respective JSON objects. Our schema organises all relevant data about the claim-verbalisation pair in a single JSON object.

WDV is a 3 star dataset according to the 5 star deployment scheme for Linked Data.[7] It is available on the web in a structured, machine-readable, and non-proprietary format. Making it 4 star by converting it into RDF is our immediate next step. Wikidata already has a well-documented RDF representation schema,[8] reified based on n-ary relationships [7]. We will make use of this schema to express the data about the claim and its components (e.g. ids, rank, labels, descriptions, values, etc.), as they are already explicitly supported by it, and it is an effective way to represent Wikidata in RDF [17]. We will then complement it with custom vocabulary in order to express the verbalisations and their crowdsourced annotations. We can do this by linking the statements, expressed in Wikidata's RDF schema as nodes, to a verbalisation node through a `wdv:verbalisation` predicate, which then is linked to its crowdsourced annotations through fitting predicates, e.g. `wdv:fluencyScore` and `wdv:adequacyScore`. We can also reuse existing vocabularies, such as LIME [9]).

Table 2 shows a breakdown of WDV. In the first column, we can identify the SEEN and UNSEEN partitions from WebNLG, as well as our added WD_UNSEEN partition built from other Wikidata classes. The second column divides

[7] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data.
[8] https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format.

```
{   "claim_id": "Q55425899$D1CB6CEC-33E4-41DF-9244-3277C2BE1FA5"
    "rank" : "normal",
    "subject_id" : "Q55425899",
    "property_id" : "P6216",
    "subject_label" : "Spring in Jølster",
    "property_label" : "copyright status",
    "object_label" : "public domain",
    "subject_desc" : "painting by Nikolai Astrup",
    "property_desc" : "copyright status for intellectual creations like
    ↪  works of art, publications, software, etc.",
    "object_desc" : "works that are no longer in copyright term or were
    ↪  never protected by copyright law",
    "subject_alias" : "no-alias",
    "property_alias" : ["copyright restriction"],
    "object_alias" : ["PD", "out of copyright", "DP"],
    "object_datatype" : "wikibase-item",
    "object" : { "value": {"entity-type": "item", "numeric-id": 19652,
    ↪  "id": 'Q19652'}, "type": "wikibase-entityid" },
    "theme_root_class_id" : "Q3305213",
    "theme_label" : "Painting",
    "verbalisation" : "Spring in J <unk> lster is in the public domain.",
    "verbalisation_unk_replaced" : "Spring in Jølster is in the public
    ↪  domain.",
    "sampling_weight" : 3538.615384615385,
    "annotations": { "fluency_scores" : [5, 4, 4, 2, 1],
                     "fluency_mean" : 3.2,
                     "fluency_median" : 4.0,
                     "adequacy_scores" : [0, 0, 1, 0, 0],
                     "adequacy_majority_voted" : 0,
                     "adequacy_percentage" : 0.8 }
}
```


Fig. 2. Example of an annotated record from WDV in JSON format

them into component themes (or pools of claims). For each theme, it then shows the number of unique properties (predicates), unique claims (calculated as N_t, as described in Sect. 3.1), and how many were annotated.

3.5 Crowd Data and Risk Analysis

Crowdsourced annotations were aggregated and added to WDV as attributes, as depicted in Sect. 3.4. In this section, we analyse these aggregated annotations and draw conclusions on the quality and reliability of WDV.

Aggregation and Reliability. Fluency scores were aggregated by calculating both median and mean, in case more or less weight, respectively, needs to be given to workers who disagree greatly with their peers. Adequacy was aggregated by majority voting, and also by calculating the percentage of workers that voted *Yes*, which we call *adequacy percentage*.

Table 2. Total number of unique properties, unique claims, and annotated claims, per partition and themes in WDV.

Partition	Theme	Properties	Claims	Annotated claims
WebNLG_SEEN	Airport	27	382	76
	Astronaut	57	351	71
	Building	67	385	63
	City	72	383	73
	ComicsCharacter	79	376	76
	Food	64	368	67
	Monument	62	380	51
	SportsTeam	49	383	75
	University	62	378	75
	WrittenWork	21	385	66
WebNLG_UNSEEN	Artist	65	384	78
	Athlete	53	385	80
	CelestialBody	25	385	83
	MeanOfTransportation	58	376	71
	Politician	56	385	75
WD_UNSEEN	ChemicalCompound	33	383	81
	Mountain	23	380	69
	Painting	29	385	50
	Street	21	384	66
	Taxon	27	385	80
ALL	ALL	439	7607	1426

Fluency has been fair to very high in most verbalisations. A fluency score of 3 indicates "Comprehensible text with minor grammatical errors", and over 96% of verbalisations find themselves with median fluency equal to or above 3. This shows our verbalisation model produces fluent text from Wikidata triples. The model also maintains very well the meaning of Wikidata claims after verbalising. Almost 93% of verbalisations are majority-voted as adequate.

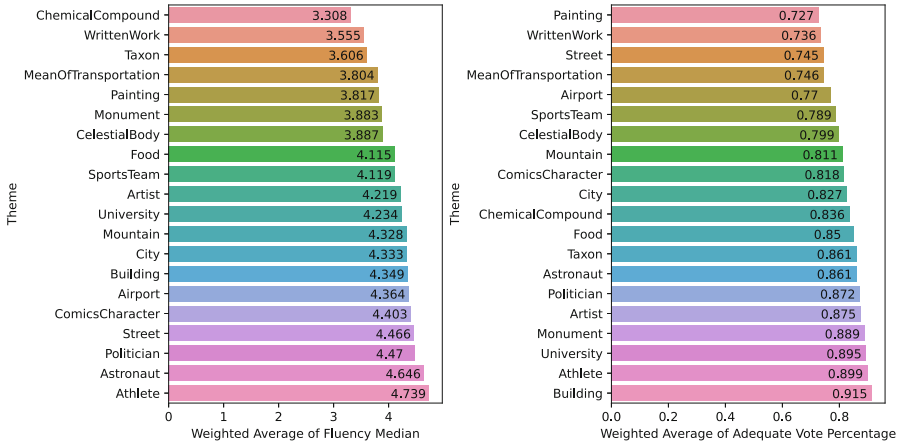
The reliability of aggregated crowdsourced data can be indicated by statistical inter-annotator agreement metrics [26] such as Krippendorff's Alpha [16]. The alpha measured for the fluency scores is 0.4272, and for the adequacy scores it is 0.4583; both indicate moderate agreement, according to the interpretations recommended by Landis & Koch [21].

Variations in Scores and Agreement. Next, we see how fluency, adequacy, and agreement might vary across the partitions and themes shown in Table 2.

We can calculate fluency and adequacy scores for each theme by making use of the sampling weights, accounting for any bias introduced by stratification. Figure 3a shows the adjusted median fluency per theme: all have from fair (above 3) to excellent (above 4) fluency, with complex and scientific themes in the lower

half. Figure 3b shows the adjusted adequacy percentage per theme, ranging from 85.7% to 99.8%.

For a bigger-picture view, we calculate the average aggregated fluency and adequacy per partition. This does not consider the sampling weights, as they are not translatable across differently stratified populations. In all aggregated metrics (i.e. mean fluency, median fluency, adequacy percentage, and majority-voted adequacy) WebNLG_SEEN performs the best, followed by WebNLG_UNSEEN, and then WD_UNSEEN. Exact metrics can be seen in Table 3. This is in line with how the model was trained and validated. However, the differences are small, signalling excellent generalisation to themes unseen both in training and validation, and also whose provenance is from an entirely different KG.



(a) (b)

Fig. 3. Median fluency (a) and adequacy percentage (b) per theme after adjusting for stratification by considering sampling weights.

Table 3. Aggregated scores and agreement per partition. Mean fluency, median fluency and adequacy percentage were averaged. Majority-Adequate Perc. is the percentage of claims whose majority-voted adequacy score was *Yes*.

	WD_UNSEEN	WebNLG_UNSEEN	WebNLG_SEEN
Mean fluency	3.684	3.884	3.91
Median fluency	3.848	4.103	4.148
Adequacy percentage	80.3%	80.6%	82%
Majority-adequate perc.	92.5%	92.8%	93.1%
Fluency scores agreement	0.466761	0.508015	0.496089
Adequacy scores agreement	0.659174	0.649527	0.654175

We calculate the agreement for each theme and partition. All themes show agreement above 0.4 on the fluency task, and above 0.6 on the adequacy task.

Fluency and adequacy agreement metrics per theme have a substantial correlation (0.63 Pearson correlation). Agreement did not vary substantially between partitions (see Table 3), showing that whether or not the model was trained or validated on a partition did not impact the workers' abilities to judge it.

4 Verbalisation Model

Our dataset relies on a pre-trained and fine-tuned data verbalisation model for its text alignment. In this section, we describe the model we have chosen and all reasons for it, as well as its training approach and hyperparameters used. We finish by evaluating its fitness for use with examples from our dataset.

4.1 Approach, Training, and Validation

Many state-of-the-art KG data verbalisation models take the graph structure into consideration [10,35,44]. GTR-LSTM [35] and DualEnc [44] both encode the graph by combining graph neural networks and recurrent sequential neural networks. Working with single-claim data, we do not need to maintain the graph's structure. Large pre-trained language models have achieved state-of-the-art results when fine-tuned and evaluated on WebNLG [14,15,32], mainly the T5 [29]. They can disregard most structure and can be applied to one or many claims at a time. Hence, we utilise the T5 (base version) as our verbalisation model, following training and evaluation methods from these works.

The T5 converts input text into output text based on a given task, such as summarisation, specified through natural language as a prefix to the input. It can also learn new tasks by being fine-tuned with new data and a new prefix [29]. Our model has been fine-tuned on WebNLG [12]. The SEEN partition is used for both training and validation/testing, while the UNSEEN partition is kept for testing only. We follow the training setup from Ribeiro et al. [32] by specifying a new prefix "translate from Graph to Text" and adding three new tokens ($\langle H \rangle$, $\langle R \rangle$, and $\langle T \rangle$) that precede the claim's subject, predicate, and object, respectively.

Each entry in the training data consists of a set aligning multiple triples to multiple sentences. We train the model by concatenating all triples in the set in a random order, marked with the new tokens, and choosing one of the verbalisations at random (as long as they were tagged as *good* by WebNLG).

Some of the hyperparameters used in the model were: a 3e-5 learning rate on a linear scheduler with no warmup steps, an Adam [19] optimiser with no weight decay and 1e-8 epsilon, a cross-entropy loss function, 100 epochs of training with early stopping patience of 15, batch size of four for both training and evaluation, and text generation was done via beam search with 3 beams. The T5 was obtained from HuggingFace [39] and trained with PyTorch Lightning [8]. A full list of hyperparameters is available in this paper's GitHub repository.

We benchmarked our model in two ways. Firstly, by evaluating aggregated crowdsourced metrics, covered in Sect. 3.5. Secondly, by calculating the BLEU score [27] for both the SEEN and UNSEEN partitions to make sure our

model matches current state-of-the-art [32]. BLEU scores were calculated with Moses [20]. The model achieved 65.51 on SEEN, 51.71 on UNSEEN, and 59.41 on ALL. We refer to Ribeiro et al. [32] for a comparison with results from other methods.

4.2 Evaluation and Output Examples

Comparisons between the model employed by us (T5-base), other language models, and other mentioned approaches to verbalising the WebNLG dataset have been carried by other authors [2,32] and are out of the scope of our paper, as well as an analysis of language variability and complexity.

To evaluate our model's fitness for use, we will now look at cases in which it performed well and poorly for fluency or adequacy. This allows us to understand how the model could improve, and in which situations its use is and is not effective. Table 4 shows a few examples of the model's inputs and outputs.

Table 4. Examples of claims verbalised by the model. From left to right, we see: the claims, their components' labels, the verbalisations, and their aggregated metrics of fluency (median) and adequacy (adequacy percentage).

Subject	Predicate	Object	Verbalisation	Flu.	Ade.
Eve Towards the Light	Height	268 centimetre	Eve Towards the Light is 268 centimetres high	2.0	1.0
Bydgoszcz	Flag	Flag of Bydgoszcz	The flag of Bydgoszcz is the flag of Bydgoszcz	0.0	1.0
Marawi	Köppen climate classification	Tropical climate	City of the Philippines and capital of the province of Lanao del Sur	4.0	0.4
Decius	Child	Hostilian	Decius is a child of Hostilian	4.0	0.4
UCAC2 28288086	Surface gravity	9 centimetre per square second	UCAC2 28288086 has a surface gravity of 9 centimetres per square second	4.0	1.0
Casa Batlló	Fabrication method	Trencadís	Casa Batlló is made using the Trencadís method	5.0	0.8

We consider a low fluency score to be under 3, when grammatical errors are not minor or text is not comprehensible. Out of over 1.4k annotated claim-verbalisation pairs, 55 had low fluency. A considerable amount of them (41%) suffer due to subject or object labels having complex syntaxes, such as IUPAC chemical nomenclatures, names of astronomical bodies, and full titles of scientific papers. These are challenging both for the model and for workers with no context or knowledge of how to use these names in a sentence. This potential misinterpretation is evidenced by 38% of all low-fluency verbalisations being simply misinterpreted by the crowd; the sentences are fluent, but have non-trivial or non-English terms that throw workers off e.g. "Eve Towards the Light is 268 centimetres high", which describes a painting. Around a third (32%) of cases

were the model's fault, either by failure to structure the predicate or by corrupting or inverting subject and object labels. However, 21% of cases could be solved by improving predicates and entity labels, or rethinking how information is stored in Wikidata; some predicates are vague or depend on qualifiers to make complete sense e.g. `inception` and `different from`, and some claims have redundant data e.g. "The flag of Bydgoszcz is the flag of Bydgoszcz".

Low adequacy is when the majority-voted option for adequacy was *No*. This corresponds to 78 verbalisations. Almost half (46.15%) consists of claims either for which the model could not properly structure the predicate e.g. "Köppen climate classification" or for which subject and predicate had complex or non-English labels. Over a third (38.4%) of these were adequate claims that were misunderstood by the crowd e.g. "Craig-y-llyn is designated as a Site of Special Scientific Interest". Somewhat often (17.9%), vague predicates and badly written labels were also to blame. Lastly, the model would sometimes (11.5%) either shift subject with object, infer information not seen in the claim (delusions), or translate words between English and German (one of T5's other learned tasks).

These cases show us that the verbalisation model can be improved either by design or through data curation. For instance, predicates that rely on qualifiers can have that information communicated to the model if the model can properly link them to specific components of the claim. We can avoid inversion of subject and object by adding direction either on the predicate labels (e.g. *child* to *has child*) or through the model's encoding. We managed to help the model understand certain predicates and entities by using alternative labels (e.g. *conflict* to *participated in conflict*), but which aliases to use is very context dependent.

Some issues are less trivial to address. Entities with syntactically complex labels hardly have simpler aliases. Vague predicates might be solved by using aliases, but this is extremely context-sensitive, and there might be good reasons why these predicates unite multiple senses under a common abstraction (e.g. `facet of` and `inception`). Finally, redundant information can emerge from Wikidata's predicates. For instance, an entity exists for the city of Bydgoszcz, and another for its flag, containing information such as its appearance. They are linked by the `flag` predicate. This makes ontological sense, but no verbal sense, as one would express this relationship as either "Bydgoszcz has a flag" or "Bydgoszcz's flag is Bydgoszcz's flag"; this is either redundant or inadequate.

5 Addressing Review Criteria

Here, we further strengthen the argument that the resources presented are not only of interest to Semantic Web researchers, but have a provable claim to adoption by them and the Wikidata research community. These resources support a line of research by the same authors on the quality of Wikidata references, which proposes crowdsourcing and computational methods to assess different dimensions of reference quality. The first part of the work assessed reference accessibility, relevance and authoritativeness based on features that are not directly related to the content of the reference themselves. It has been recently awarded

the Wikimedia Research Paper of the Year 2022, from among 230 peer-reviewed papers. The judges highlighted the importance of the research problem (reference quality) and the value of the solution proposed, especially in a multilingual setting. WDV directly builds on top of this, by feeding into computational methods that allow us to assess reference quality also in terms of the actual content in the reference source. It has already made possible the authors' efforts toward automated fact verification in Wikidata.

Wikidata recognises references as essential in its own guidelines, stating that "Wikidata is not a database that stores facts about the world, but a secondary knowledge base that collects and links to references to such knowledge".[9] They promote reference quality assurance efforts, as many open phabricator tickets show.[10,11] The Wikidata editing community also discusses at length the need for automated techniques for reference quality assessment.[12,13]

6 Conclusion

In this paper, we have presented WDV: a large dataset for the verbalisation of single triple-based claims from Wikidata (a collaborative KG). It directly aligns claims to natural language sentences that aim at being grammatically well-written and transmitting the same meaning. WDV was created to provide a data-to-text resource that covers a wide range of entities, topics, and predicates in Wikidata. More importantly, it does so in a balanced manner, so that specific themes are not overly represented. We also presented and carried an evaluation workflow of the fluency and adequacy of its natural language sentences, concluding that they have very high levels of both metrics.

We believe this dataset constitutes a valuable step towards understanding how to efficiently carry the verbalisation of triple claims from Wikidata and KGs in general. Bridging the gap between labelled triple components and natural language is crucial to implementing downstream NLP tasks in the KG. One such task that can be helped immensely by this resource is the automated fact-checking of KG claims based on the textual information found in the references they cite. Finally, WDV, alongside the annotation workflow we have defined, can promote the evaluation, through a human perspective, of NLG models performances without relying on algorithmic metrics. WDV's construction process can also be extended to include languages other than English by using multilingual LMs and training data akin to WebNLG.

Acknowledgements. This research received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 812997.

[9] https://www.wikidata.org/wiki/Help:Statements.
[10] https://phabricator.wikimedia.org/T90881.
[11] https://phabricator.wikimedia.org/T156389.
[12] https://www.wikidata.org/wiki/Property_talk:P1456.
[13] https://www.wikidata.org/wiki/Wikidata:Project_chat/Archive/2017/10#Proposal_on_citation_overkill.

References

1. Amaral, G., Piscopo, A., Kaffee, L.A., Rodrigues, O., Simperl, E.: Assessing the Quality of sources in Wikidata across languages: a hybrid approach. J. Data Inf. Qual. **13**(4) (2021). https://doi.org/10.1145/3484828
2. Bosc, T., Cabrio, E., Villata, S.: DART: a dataset of arguments and their relations on Twitter. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1258–1263. European Language Resources Association (ELRA), Portorož, Slovenia, May 2016
3. Burnham, M.J., Le, Y.K., Piedmont, R.L.: Who is Mturk? Personal characteristics and sample consistency of these online workers. Mental Health Relig. Cult. **21**(9–10), 934–944 (2018). https://doi.org/10.1080/13674676.2018.1486394
4. Difallah, D., Filatova, E., Ipeirotis, P.: Demographics and dynamics of mechanical turk workers. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 135–143. WSDM 2018. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159661
5. Ellis, J., Getman, J., Graff, D., Strassel, S.: TAC KBP Comprehensive English Source Corpora 2009–2014 (2018). https://doi.org/11272.1/AB2/VC89SM
6. Elsahar, H., et al.: T-REx: a large scale alignment of natural language with knowledge base triples. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan, May 2018
7. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the linked data web. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 50–65. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_4
8. Falcon, W.: The PyTorch Lightning team: PyTorch Lightning, March 2019. https://doi.org/10.5281/zenodo.3828935, https://github.com/Lightning-AI/lightning
9. Fiorelli, M., Stellato, A., McCrae, J.P., Cimiano, P., Pazienza, M.T.: LIME: The metadata module for OntoLex. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.) The Semantic Web. Latest Advances and New Domains, pp. 321–336. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-18818-8_20
10. Gao, H., Wu, L., Hu, P., Xu, F.: RDF-to-Text generation with graph-augmented structural neural encoders. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pp. 3030–3036. International Joint Conferences on Artificial Intelligence Organization, July 2020. https://doi.org/10.24963/ijcai.2020/419, main track
11. Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L.: Creating training corpora for NLG micro-planners. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 179–188. Association for Computational Linguistics, Vancouver, Canada, July 2017. https://doi.org/10.18653/v1/P17-1017
12. Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L.: Creating training corpora for NLG micro-planners. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 179–188. Association for Computational Linguistics, Vancouver, Canada, July 2017. https://doi.org/10.18653/v1/P17-1017

13. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., au2, H.D.I., Crawford, K.: Datasheets for Datasets (2020)
14. Guo, Q., et al.: P^2: a plan-and-pretrain approach for knowledge graph-to-text generation. In: Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), pp. 100–106. Association for Computational Linguistics, Dublin, Ireland (Virtual), December 2020. https://aclanthology.org/2020.webnlg-1.10
15. Harkous, H., Groves, I., Saffari, A.: Have your text and use it too! End-to-End neural data-to-text generation with semantic fidelity. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 2410–2424. International Committee on Computational Linguistics, Barcelona, Spain (Online), December 2020. https://doi.org/10.18653/v1/2020.coling-main.218
16. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. **1**(1), 77–89 (2007). https://doi.org/10.1080/19312450709336664
17. Hernández, D., Hogan, A., Krötzsch, M.: Reifying RDF: What works well with Wikidata? In: SSWS@ISWC (2015)
18. Huff, C., Tingley, D.: "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. Res. Polit. **2**(3) (2015). https://doi.org/10.1177/2053168015604648
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2015)
20. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180. Association for Computational Linguistics, Prague, Czech Republic, June 2007
21. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977). http://www.jstor.org/stable/2529310
22. van der Lee, C., Gatt, A., van Miltenburg, E., Krahmer, E.: Human evaluation of automatically generated text: current trends and best practice guidelines. Comput. Speech Lang. **67**, 101151 (2021). https://doi.org/10.1016/j.csl.2020.101151
23. Ma, K., Cheng, H., Liu, X., Nyberg, E., Gao, J.: Open Domain Question Answering over Virtual Documents: A Unified Approach for Data and Text (2021)
24. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011. Association for Computational Linguistics, Suntec, Singapore, August 2009. https://aclanthology.org/P09-1113
25. Novikova, J., Dušek, O., Curry, A.C., Rieser, V.: Why we need new evaluation metrics for NLG. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2241–2252. Association for Computational Linguistics, Copenhagen, Denmark, September 2017. https://doi.org/10.18653/v1/D17-1238
26. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the International Conference on Multimedia Information Retrieval, pp. 557–566. MIR 2010. Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1743384.1743478

27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. ACL 2002 (2002). https://doi.org/10.3115/1073083.1073135
28. Tanon, T.P., Vrandečić, D., Schaffert, S., Steiner, T., Pintscher, L.: From freebase to Wikidata: the great migration. In: Proceedings of the 25th International Conference on World Wide Web, pp. 1419–1428. WWW 2016, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2016). https://doi.org/10.1145/2872427.2874809
29. Raffel, C., et al.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. CoRR abs/1910.10683 (2019). http://arxiv.org/abs/1910.10683
30. Ramírez, J., et al.: On the state of reporting in crowdsourcing experiments and a checklist to aid current practices. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (2021). https://doi.org/10.1145/3479531
31. Reiter, E.: A structured review of the validity of BLEU. Comput. Linguist. **44**(3), 393–401 (2018). https://doi.org/10.1162/coli_a_00322
32. Ribeiro, L.F.R., Schmitt, M., Schütze, H., Gurevych, I.: Investigating pretrained language models for graph-to-text generation. In: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pp. 211–227. Association for Computational Linguistics, Online, November 2021. https://aclanthology.org/2021.nlp4convai-1.20
33. Sulem, E., Abend, O., Rappoport, A.: BLEU is not suitable for the evaluation of text simplification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 738–744. Association for Computational Linguistics, Brussels, Belgium, October 2018. https://doi.org/10.18653/v1/D18-1081
34. Toutanova, K., Chen, D.: Observed versus latent features for knowledge base and text inference (2015)
35. Trisedya, B.D., Qi, J., Zhang, R., Wang, W.: GTR-LSTM: a triple encoder for sentence generation from RDF data. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1627–1637. Association for Computational Linguistics, Melbourne, Australia, July 2018. https://doi.org/10.18653/v1/P18-1151
36. Vlachos, A., Riedel, S.: Identification and verification of simple claims about statistical properties. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2596–2601. Association for Computational Linguistics, Lisbon, Portugal, September 2015. https://doi.org/10.18653/v1/D15-1312
37. Vougiouklis, P., et al.: Neural Wikipedian: generating textual summaries from knowledge base triples. J. Web Semant. **52–53**, 1–15 (2018). https://doi.org/10.1016/j.websem.2018.07.002
38. Vougiouklis, P., Maddalena, E., Hare, J., Simperl, E.: Point at the triple: generation of text summaries from knowledge base triples (extended abstract). In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pp. 5080–5084. International Joint Conferences on Artificial Intelligence Organization, July 2020. https://doi.org/10.24963/ijcai.2020/711, journal track
39. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online, October 2020

40. Yang, X., Nie, F., Feng, Y., Liu, Q., Chen, Z., Zhu, X.: Program Enhanced fact verification with verbalization and graph attention network. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7810–7825. Association for Computational Linguistics, Online, November 2020. https://doi.org/10.18653/v1/2020.emnlp-main.628
41. Yang, X., Zhu, X.: Exploring Decomposition for Table-based Fact Verification (2021)
42. Yao, L., Haghighi, A., Riedel, S., McCallum, A.: Structured relation discovery using generative models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1456–1466. Association for Computational Linguistics, Edinburgh, Scotland, UK, July 2011. https://aclanthology.org/D11-1135
43. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 35–45. Association for Computational Linguistics, Copenhagen, Denmark, September 2017. https://doi.org/10.18653/v1/D17-1004
44. Zhao, C., Walker, M., Chaturvedi, S.: Bridging the structural gap between encoding and decoding for data-to-text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2481–2491. Association for Computational Linguistics, Online, July 2020. https://doi.org/10.18653/v1/2020.acl-main.224