

Towards More Informative List Verbalisations

Lea Krause^{1,*}, Pia Sommerauer¹ and Piek Vossen¹

¹Vrije Universiteit Amsterdam, The Netherlands

Abstract

In this paper we propose the task of list verbalisation within a Knowledge Graph Question Answering system. Inspired by the Gricean Maxims of Quantity, Relation, and Manner we show a proof of concept ranking answer candidates through graph-based and language model-based measurements for on the one hand popularity and on the other hand a more pragmatically informed context. Our findings show that in our current set-up graph-based measures work best, while language model-based systems need further refinement and may benefit from approaches such as fine-tuning or prompting. We evaluate our approach with a user study and give insights into promising future directions of the task.

Keywords

KGQA, List verbalisation, Ranking, Summarisation, Gricean maxims

1. Introduction

Question Answering (QA) systems are becoming more prevalent in both research and real-world applications such as virtual assistants like Siri or Alexa. With vast amounts of structured knowledge available in ontologies, they have been a key element in furthering QA development and adaptation. Up until recently their answers were however limited to formal query responses, which limit usefulness for conversational systems as well as non-expert users. To increase naturalness and understanding [1], a recent focus of Knowledge Graph Question Answering (KGQA) systems has been the verbalisation of the query answer [1, 2, 3, 4]. They take the generated formal query response and present it in natural language, for example by taking into account the wording of the question [3]. While investigating existing data sets we noticed that (long) list answers are currently poorly dealt with or excluded from the data. We see filling this gap as a natural next step in towards more informative and natural KGQA verbalisation.

Responses consisting of a whole list present the user with information overload, in particular when questions result in a large number of answers. They will most likely not succeed in finding an answer that is informative to them. From a pragmatic perspective, we can analyse the likely success of a question-answer exchange through the lens of the co-operative principle consisting of four maxims proposed by Grice [5]. We focus on the maxims of Quantity, Relation, and Manner to improve the communication of the results. Consequently, we slightly alter the

KGSum2022: International Workshop on Knowledge Graph Summarization, October 23, 2022, Hangzhou, China


*Corresponding author.

✉ l.krause@vu.nl (L. Krause); pia.sommerauer@vu.nl (P. Sommerauer); p.t.j.m.vossen@vu.nl (P. Vossen)

🌐 <https://lkra.github.io/> (L. Krause); <https://piasommerauer.github.io> (P. Sommerauer); <https://vossen.info> (P. Vossen)

🆔 0000-0001-7187-5224 (L. Krause); 0000-0003-3593-1465 (P. Sommerauer); 0000-0002-6238-5941 (P. Vossen)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

suggested verbalisation template, cut down the list of given answers and pick the most relevant ones. This can be done in a generic way, picking the most popular one, or in a context-dependent way, in which we consider previous conversations, the user, or the specific application. We give preliminary results of our method based on a small user study.

Our approach combines summarisation and ranking, to find the most informative answer for the user. It is a first step to, on the one hand, the verbalisation of more complex formal query responses, and on the other hand the inclusion of context in the summarisation and verbalisation of structured data.

Our contributions in this paper are as follows:

1. **Task** We create the new task of list verbalisation, which was previously neglected (VQuAnDa [2]) or excluded (VANiLLa [3]) from KG verbalisation tasks. The focus lies on the combination with context cues as secondary properties which can personalise or diversify the verbalisation.
2. **Implementation** We show a proof of concept and implement first measurements for popularity and context cue specific verbalisation.
3. **System Comparison** We conduct a system comparison between language model-based and graph-based metrics and verify our set-ups with human evaluation.

2. Related work

As our task and evaluation approach are influenced by a range of research fields, we will describe previous approaches from a range of fields including KGQA systems, data-to-text generation, summarisation, ranking, and their evaluations. To the best of our knowledge, this paper is the first exploration of list verbalisation in the context of Gricean maxims.

2.1. QA tasks and systems

Most QA systems can be divided into two groups. The first group works with unstructured data such as SQuAD [6, 7] or CoQA [8], while the second is based on structured knowledge such as DBpedia [9] or Wikidata [10].

In a KGQA task the question is given in natural language, gets translated into a query, queries a database and retrieves an answer. The most common data in KGQA sets are the editions of QALD [11] and LC-QuAD [12]. The former’s focus has over the years shifted towards multi-linguality [13]. The latter provides a data set of 5000 complex question-query pairs over DBpedia.

2.2. A data to text problem

Our focus lies on the verbalisation of the formal query response, turning the problem into a data-to-text task. Most approaches for this task are based on WebNLG [14] and more recently AGENDA [15] and use either transformers [15, 16] or Graph Neural Nets (GNN) [17]. Most of the previously described tasks are however too domain specific to be applied to an open-domain knowledge graph like DBpedia or Wikidata.

Data sets that have specifically explored verbalisation in a KGQA context are VQuAnDa [2], VANiLLa [3], ParaQA [1]. VQuAnDa [2] consists of 5000 complex questions, SPARQL queries and answer verbalisation. ParaQA [1] expands the verbalised answers of VQuAnDa with two to eight paraphrased natural language responses per question. VANiLLa [3] contains 100k simple questions with their respective queries and verbalised answers, adapted from CSQA [18] and SimpleQuestionsWikidata [19]. When dealing with lists, VQuAnDa and ParaQA give a filler token [answer] instead of the full list in their verbalised sentences. VANiLLa, on the other hand, excludes list question-answer pairs. We focus on the question-answer pairs containing lists in VQuAnDa.

List verbalisation can be approached from the perspective of ranking with respect to relevance. Approaches considering the ranking of multiple query response options have been extensively studied in the field Information Retrieval, ranking from early symbolic approaches [20] to recent neural ranking models [21, 22]. Within graphical entity summarisation there have been approaches focusing on relevance-oriented entity summarisation [23] and diversity [24] though they differ from ours by specifically aiming to provide generic summaries instead of a context dependent ones.

Context in Information Retrieval and Summarisation systems is mostly based on personalisation of the results for the user based on their previous interaction with the system [25, 26, 27, 28]. Other forms of context can include location metadata or time [29].

2.3. Evaluation in terms of Gricean Maxims

Our evaluation is guided by the Gricean maxims [5] of *Quantity*, *Relation* and *Manner*. They define principles that should be fulfilled in order to successfully communicate. The maxim of *Quantity* states that all necessary information should be given, but not more than the intent of the question requires. This guides our general reduction of answers given and reformulation of the answer template. The maxim of *Relation* stresses the importance of adapting the information to be relevant to the communication partner, in this case the user. The maxim of *Manner* states that communication should be orderly and brief, avoiding obscurity and ambiguity. The maxims are widely used within pragmatics and have been applied to time series summarisation [30] and recently chatbots [31]. Other common metrics in NLG evaluation are ROUGE, BLEU and METEOR; these are, however, based on n-gram similarity which is not suitable for our case since there is no ground truth available for comparison. Furthermore, these measures fail to take pragmatic factors like purpose or context into account [32]. We aim to go beyond existing approaches and directly evaluate the output of our systems through human judgements of whether the verbalisations adhere to the maxims of *Quantity*, *Relation*, and *Manner*.

3. Task

We propose a the new task of list verbalisation with respect to the Gricean maxims of *Quantity*, *Relation*, and *Manner*. Given a long answer list, the task is to select a small number of items that serve to illustrate the most relevant information conveyed by the list.

What is relevant to the user can be highly situation-dependent and depend on individual factors specific to the user. In this first approach, we define a number of specific **contextual**

cues that define what is relevant for the user. In the simplest scenario, we assume that the user is simply interested in the most popular examples from the entire list.

Another reason to not only work with the highest scores is that they enforce popularity bias. DBpedia is based on Wikipedia which is predominantly white, western and male [33, 34, 35, 36]. Cues are an option to diversify or personalise results and increase the exposure of non-popular items.

The contextual cues should reflect relatively open secondary factors. We currently use the following: location (place of birth, place of publication, nationality), time (recency, time of publication, date of birth), genre, and gender.

We illustrate the task using the following question / verbalisation pair:

Question: "Whose work is in the Musée d'Orsay?"

VQuAnDa answer template: "The artists of the artworks located in the Musée d'Orsay are [answer]."

The place holder [answer] is filled with 30 names (see step 1 Figure 1), which is still on the lower end, as the number of items in such lists can exceed 1000. Giving an answer consisting of 30 names would be a violation of the Gricean Maxim of Quantity since it presents more information than can be assumed was intent of the user's question. This can lead to an information overload and the relevant information being buried. An answer consisting of three very popular artists (Vincent van Gogh, Paul Gauguin, Henri Matisse), however, would not violate the maxim and can be considered relevant. If we want to consider a particular context (e.g. a previous conversational context of female artists), another subset of artists can provide a relevant answer (e.g. examples of female artists).

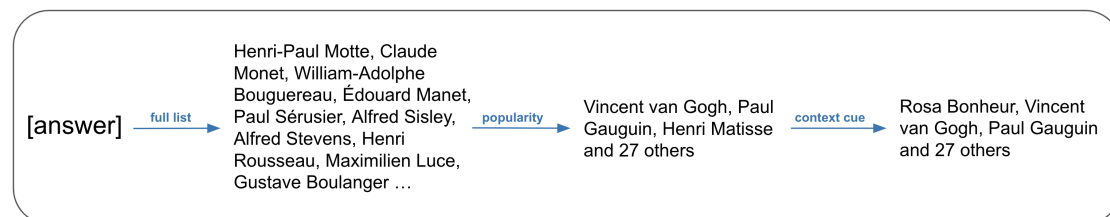


Figure 1: An example of how our model deals with and modifies list answers for the question "Whose work is in the Musée d'Orsay?". The popularity metric used is density and the context cue is *female artists*.

4. Method

We explore two approaches to list verbalisation: A language-model based approach and a knowledge-graph based approach. In both approaches, we aim to rank the answer list with respect to popularity or a particular contextual cue to retrieve the most relevant answers. After a brief outline of the data we work with (Section 4.1), we introduce the two ranking approaches we use to generate answers (Section 4.2). Finally, we describe the set-up of our human evaluation (Section 4.3).

4.1. Data

As described in section 2, available verbalisation data sets and systems do not currently deal with long list answers in an adequate manner. VQuAnDa [2] contains 444 examples (8.9%) that have 15 or more answers and are therefore not verbalised. From these examples, we select ten instances for our exploratory experiments. The lists range in number of answer candidates from 13 to 1239. We only select 10 instances, as our evaluations relies on human judgements of system output.

4.2. Ranking and answer-generation

We construct list verbalisations containing the top- k list items retrieved by our raking approaches. The final verbalisation is presented through a template. Consider the example below showing the answer to the questions about artists in the Musée d’Orsay (introduced in Section 3).

New answer template: "The artists of the artworks located in the Musée d’Orsay are for example [top- k] and [n - top- k] others."

Our template is only going to verbalise the top- k answers and enumerates the rest of the list, to still convey the information of more answers being present.

We use two different techniques to identify the top- k answer candidates, the probability given to an answer candidate by a language model (GPT-2 [37], LM) and the density of a constructed graph for an answer candidate. Both ranking approaches only consider the items already provided in the list.

4.2.1. Language model-based measures

We can expect that language model prediction can, at least to some degree, reflect the relevance of an answer-candidate with respect to a particular context. To score the answer options, we first create a template to fill with singular answers instead of plural. This is done by converting the subject and predicate of the sentence into singular. For the museum example this means:

Singular answer template: "An artist of the artworks located in the Musée d’Orsay is [answer]."

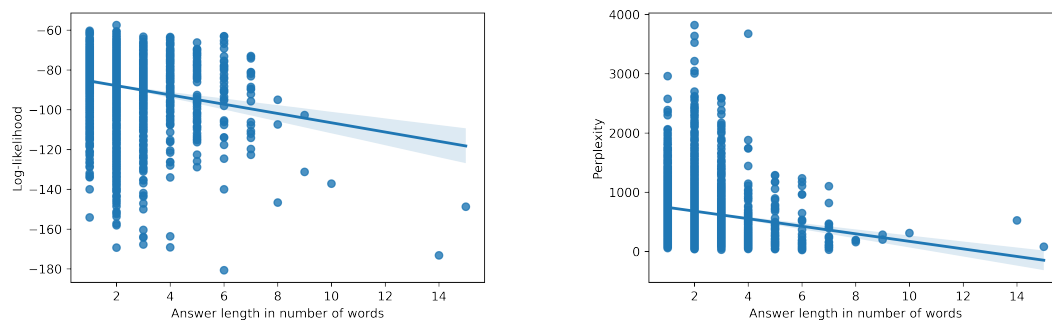
When ranking answers with respect to a contextual cue, we currently add the cue to the template sentence, after which the answer candidates are ranked:

Singular answer template with gender cue: "A *female* artist of the artworks located in the Musée d’Orsay is [answer]."

We fill the answer slot for each answer candidate and use GPT-2¹ to calculate the answer’s log-likelihood and perplexity. Upon initial qualitative exploration, we noticed that log-likelihood

¹<https://github.com/simonepri/lm-scorer>

provided better answers than expected, despite its tendency to penalise longer answers (e.g. names consisting of multiple tokens). As expected, perplexity also resulted in high-quality answers. When investigating the correlation between answer length and LM scoring more closely we found that while maximising for log-likelihood significantly penalises answers for being long, minimising for perplexity overly rewards them (see Figure 2a). We include both metrics in our experiments to establish whether humans may have a preference for shorter or longer list items in the answers.



(a) Influence of answer length on log-likelihood

(b) Influence of answer length on perplexity

Figure 2: Comparison of log-likelihood and perplexity as affected by the answer options length in words. For log-likelihood the answer length is negatively correlated, meaning shorter answers were preferred. Contrastively, perplexity goes down the longer the answer length improving the score, meaning longer answers were preferred.

4.2.2. Graph-based measures

To retrieve relevant answers from a graph, we construct the corresponding graph of each answer candidate containing all outgoing nodes. To identify the most relevant graph we use **density** (D):

$$D = \frac{|E|}{|N|(|N| - 1)}$$

where E is the number of edges and N is the number of nodes in the graph. Density reflects how well connected a graph is and can thus be seen as an approximation of popularity. More recently it has also been shown to correlate well with human judgements in the evaluation of open-domain dialogue systems [38]. The higher the density of the graph, the higher we rank the corresponding answer option. For an example result see step 2 in Figure 1. To calculate density for lists with specific contextual cues, we add an additional restriction containing the context cue.

While we considered a more commonly used metric such as PageRank, the implementation would have led to the loss of examples since DBpedia does not currently provide a PageRank measurement in its SPARQL endpoint. Reformating to for example the Wikidata format would have led to either loss of examples and answer options due to an imperfect mapping or labour intensive human input.

4.3. Human evaluation

The goal of our evaluation is to establish whether potential users prefer the verbalisations created by our ranking approaches over the original, long lists. In addition, we check whether the ranked examples in our verbalisations are perceived as more helpful than random examples. We thus create verbalised answer options with rankings according to randomness (unranked), log-likelihood, perplexity and density.

In order to verify our ranking we conducted a survey with 10 participants. For each of the selected examples they were presented with five verbalised answer rankings: Random order, ranked by log-likelihood, by perplexity, and by density. As a fifth alternative we included an option to indicate if they thought the full list would have been more suitable. They then ranked the options from most (1st choice) to least (5th choice) relevant. Context cues were included as a highlighted word below the question and participants were instructed to include them in their ranking. Participants were told to not rank questions if they were too unfamiliar with the topic in question².

5. Results

In our experiments we are taking the first steps in providing a proof of concept for our extractive summarisation and context cues. For each approach, we fill our answer template (see Section 4.2) with the top- $k=3$ ranking results. The results of the human ratings are summarised in Table 1 (general popularity ranking) and Table 2 (ranking with respect to context cues). The Tables show how many times an answer created through the different ranking methods was placed on a particular rank by the human participants.

Table 1

Choice ranking for popularity

Choice	1st	2nd	3rd	4th	5th
Random	6	17	32	32	2
Log-likelihood	18	29	20	21	1
Perplexity	13	16	28	30	2
Density	47	27	9	5	1
Full list	5	0	0	1	83

Table 2

Choice ranking for context cue options

Choice	1st	2nd	3rd	4th	5th
Random	13	28	25	24	0
Log-likelihood	6	20	32	32	0
Perplexity	6	24	32	28	0
Density	64	18	1	6	1
Full list	1	0	0	0	89

The participants clearly favoured density over the other measures for both popularity and context cue options. Log-likelihood outperforms perplexity in case of popularity, but both are ranked lower than random when including the context cue. A promising finding is that even for shorter lists, the adjusted verbalisation was preferred over the full list. The low performance of the LM-based measurements could be improved by fine-tuning on the task or using prompting, which we are considering as next steps.

²One participant accidentally left out a question in the popularity ranking option leading to one score less than in the context cue option.

6. Conclusion

We identified the task of list verbalisation as a subtask of KGQA Answer Verbalisation. We show a first proof of concept ranking answers incorporating both graph-based and language model-based measurements to identify the most informative answer candidates. We base our approach on the Gricean Maxims and show that popularity alone might not suffice to create an informative answer. We are currently including only the Maxims of Quantity, Relation, and Manner, but hope to expand to the Maxim of Quality in the future. We will do this by moving from our current extractive summaries to more abstractive renditions. Instead of slot filling we will work on full pipeline or end-to-end verbalisation systems.

Acknowledgments

This research was funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) through the *Hybrid Intelligence Centre* via the Zwaartekracht grant (024.004.022), and the *Spinoza* grant (SPI 63-260) awarded to Piek Vossen. We would also like to thank the reviewers for their excellent feedback that enhanced this paper. All remaining errors are our own.

References

- [1] E. Kacupaj, B. Banerjee, K. Singh, J. Lehmann, ParaQA: A Question Answering Dataset with Paraphrase Responses for Single-Turn Conversation, arXiv:2103.07771 [cs] (2021). URL: <http://arxiv.org/abs/2103.07771>, arXiv: 2103.07771.
- [2] E. Kacupaj, H. Zafar, J. Lehmann, M. Maleshkova, VQuAnDa: Verbalization Question ANSwering DATaset, in: A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, M. Cochez (Eds.), *The Semantic Web, Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2020, pp. 531–547. doi:10.1007/978-3-030-49461-2_31.
- [3] D. Biswas, M. Dubey, M. R. A. H. Rony, J. Lehmann, VANiLLa : Verbalized Answers in Natural Language at Large Scale, arXiv:2105.11407 [cs] (2021). URL: <http://arxiv.org/abs/2105.11407>, arXiv: 2105.11407 version: 1.
- [4] E. Kacupaj, S. Premnadh, K. Singh, J. Lehmann, M. Maleshkova, VOGUE: Answer Verbalization through Multi-Task Learning, arXiv:2106.13316 [cs] (2021). URL: <http://arxiv.org/abs/2106.13316>, arXiv: 2106.13316 version: 2.
- [5] H. P. Grice, *Logic and conversation*, in: *Speech acts*, Brill, 1975, pp. 41–58.
- [6] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: <https://aclanthology.org/D16-1264>. doi:10.18653/v1/D16-1264.
- [7] P. Rajpurkar, R. Jia, P. Liang, Know What You Don’t Know: Unanswerable Questions for SQuAD, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics,

- tics, Melbourne, Australia, 2018, pp. 784–789. URL: <https://aclanthology.org/P18-2124>. doi:10.18653/v1/P18-2124.
- [8] S. Reddy, D. Chen, C. D. Manning, CoQA: A Conversational Question Answering Challenge, *Transactions of the Association for Computational Linguistics* 7 (2019) 249–266. URL: <https://aclanthology.org/Q19-1016>. doi:10.1162/tac1_a_00266, place: Cambridge, MA Publisher: MIT Press.
- [9] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A Nucleus for a Web of Open Data, in: K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (Eds.), *The Semantic Web, Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2007, pp. 722–735. doi:10.1007/978-3-540-76298-0_52.
- [10] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* 57 (2014) 78–85. Publisher: ACM New York, NY, USA.
- [11] V. Lopez, C. Unger, P. Cimiano, E. Motta, Evaluating question answering over linked data, *Web Semantics Science Services And Agents On The World Wide Web* 21 (2013) 3–13. doi:10.1016/j.websem.2013.05.006, publisher: Elsevier.
- [12] P. Trivedi, G. Maheshwari, M. Dubey, J. Lehmann, LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs, in: C. d’Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin (Eds.), *The Semantic Web – ISWC 2017*, volume 10588, Springer International Publishing, Cham, 2017, pp. 210–218. URL: http://link.springer.com/10.1007/978-3-319-68204-4_22. doi:10.1007/978-3-319-68204-4_22, series Title: *Lecture Notes in Computer Science*.
- [13] A. Perevalov, D. Diefenbach, R. Usbeck, A. Both, QALD-9-plus: A Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers, in: *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, 2022, pp. 229–234. doi:10.1109/ICSC52841.2022.00045, iSSN: 2325-6516.
- [14] T. Castro Ferreira, D. Moussallem, E. Kraemer, S. Wubben, Enriching the WebNLG corpus, in: *Proceedings of the 11th International Conference on Natural Language Generation*, Association for Computational Linguistics, Tilburg University, The Netherlands, 2018, pp. 171–176. URL: <https://aclanthology.org/W18-6521>. doi:10.18653/v1/W18-6521.
- [15] R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, H. Hajishirzi, Text Generation from Knowledge Graphs with Graph Transformers, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2284–2293. URL: <https://aclanthology.org/N19-1238>. doi:10.18653/v1/N19-1238.
- [16] M. Schmitt, L. F. R. Ribeiro, P. Duffer, I. Gurevych, H. Schütze, Modeling Graph Structure via Relative Position for Text Generation from Knowledge Graphs, in: *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, Association for Computational Linguistics, Mexico City, Mexico, 2021, pp. 10–21. URL: <https://aclanthology.org/2021.textgraphs-1.2>. doi:10.18653/v1/2021.textgraphs-1.2.
- [17] D. Marcheggiani, L. Perez-Beltrachini, Deep Graph Convolutional Encoders for Structured Data to Text Generation, in: *Proceedings of the 11th International Conference on Natural Language Generation*, Association for Computational Linguistics, Tilburg University, The

- Netherlands, 2018, pp. 1–9. URL: <https://aclanthology.org/W18-6501>. doi:10.18653/v1/W18-6501.
- [18] A. Saha, V. Pahuja, M. Khapra, K. Sankaranarayanan, S. Chandar, Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph, *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (2018). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11332>. doi:10.1609/aaai.v32i1.11332, number: 1.
- [19] D. Diefenbach, T. P. Tanon, K. Singh, P. Maret, Question Answering Benchmarks for Wikidata, in: *ISWC 2017, Vienne, Austria, 2017*. URL: <https://hal.archives-ouvertes.fr/hal-01637141>.
- [20] S. Chaudhuri, G. Das, V. Hristidis, G. Weikum, Probabilistic information retrieval approach for ranking of database query results, *ACM Transactions on Database Systems* 31 (2006) 1134–1168. URL: <https://dl.acm.org/doi/10.1145/1166074.1166085>. doi:10.1145/1166074.1166085.
- [21] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, A Deep Look into neural ranking models for information retrieval, *Information Processing & Management* 57 (2020) 102067. URL: <https://www.sciencedirect.com/science/article/pii/S0306457319302390>. doi:10.1016/j.ipm.2019.102067.
- [22] Y. Shen, Y. Deng, M. Yang, Y. Li, N. Du, W. Fan, K. Lei, Knowledge-aware Attentive Neural Network for Ranking Question Answer Pairs, in: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018*, pp. 901–904. URL: <https://doi.org/10.1145/3209978.3210081>. doi:10.1145/3209978.3210081.
- [23] A. Thalhammer, N. Lasierra, A. Rettinger, LinkSUM: Using Link Analysis to Summarize Entity Data, in: A. Bozzon, P. Cudre-Maroux, C. Pautasso (Eds.), *Web Engineering, Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2016, pp. 244–261. doi:10.1007/978-3-319-38791-8_14.
- [24] M. Sydow, M. Piłkuła, R. Schenkel, The notion of diversity in graphical entity summarisation on semantic knowledge graphs, *Journal of Intelligent Information Systems* 41 (2013) 109–149. URL: <https://doi.org/10.1007/s10844-013-0239-6>. doi:10.1007/s10844-013-0239-6.
- [25] T. Safavi, C. Belth, L. Faber, D. Mottin, E. Muller, D. Koutra, Personalized Knowledge Graph Summarization: From the Cloud to Your Pocket, in: *2019 IEEE International Conference on Data Mining (ICDM), IEEE, Beijing, China, 2019*, pp. 528–537. URL: <https://ieeexplore.ieee.org/document/8970788/>. doi:10.1109/ICDM.2019.00063.
- [26] L. Faber, D. Koutra, Adaptive personalized knowledge graph summarization, in: *Proceedings of the 14th international workshop on mining and learning with graphs (MLG), 2018*.
- [27] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, L. Zettlemoyer, QuAC: Question Answering in Context, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018*, pp. 2174–2184. URL: <https://aclanthology.org/D18-1241>. doi:10.18653/v1/D18-1241.
- [28] P. Wu, Q. Zhou, Z. Lei, W. Qiu, X. Li, Template Oriented Text Summarization via Knowledge

- Graph, 2018. doi:10.1109/ICALIP.2018.8455241, pages: 83.
- [29] P. N. Bennett, F. Radlinski, R. W. White, E. Yilmaz, Inferring and using location metadata to personalize web search, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 135–144. URL: <https://doi.org/10.1145/2009916.2009938>. doi:10.1145/2009916.2009938.
- [30] S. G. Sripada, E. Reiter, J. Hunter, J. Yu, Generating English summaries of time series data using the Gricean maxims, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03, Association for Computing Machinery, New York, NY, USA, 2003, pp. 187–196. URL: <https://doi.org/10.1145/956750.956774>. doi:10.1145/956750.956774.
- [31] V. Setlur, M. Tory, How do you Converse with an Analytical Chatbot? Revisiting Gricean Maxims for Designing Analytical Conversational Behavior, in: CHI Conference on Human Factors in Computing Systems, ACM, New Orleans LA USA, 2022, pp. 1–17. URL: <https://dl.acm.org/doi/10.1145/3491102.3501972>. doi:10.1145/3491102.3501972.
- [32] E. Lloret, L. Plaza, A. Aker, The challenging task of summary evaluation: an overview, *Language Resources and Evaluation* 52 (2018) 101–148. URL: <https://doi.org/10.1007/s10579-017-9399-2>. doi:10.1007/s10579-017-9399-2.
- [33] P. Konieczny, M. Klein, Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator, *New Media & Society* 20 (2018) 4608–4633. URL: <https://doi.org/10.1177/1461444818779080>. doi:10.1177/1461444818779080, publisher: SAGE Publications.
- [34] M. Hinnosaar, Gender inequality in new media: Evidence from Wikipedia, *Journal of Economic Behavior & Organization* 163 (2019) 262–276. URL: <https://www.sciencedirect.com/science/article/pii/S0167268119301234>. doi:10.1016/j.jebo.2019.04.020.
- [35] J. M. Ezell, Empathy plasticity: decolonizing and reorganizing Wikipedia and other online spaces to address racial equity, *Ethnic and Racial Studies* 44 (2021) 1324–1336. URL: <https://www.tandfonline.com/doi/full/10.1080/01419870.2020.1851383>. doi:10.1080/01419870.2020.1851383.
- [36] L. M. Bridges, R. Pun, R. A. Arteaga (Eds.), *Wikipedia and Academic Libraries*, Michigan Publishing, 2021. URL: <https://hdl.handle.net/2027/fulcrum.cv43p013f>. doi:10.3998/mpub.11778416.
- [37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners (2019) 24.
- [38] S. Báez Santamaría, P. Vossen, T. Baier, Evaluating Agent Interactions Through Episodic Knowledge Graphs, 2022. URL: <http://arxiv.org/abs/2209.11746>. doi:10.48550/arXiv.2209.11746, arXiv:2209.11746 [cs].