Knowledge Graph Population with Out-of-KG Entities

Cedric Möller[0000-0001-6700-3482]

Semantic Systems Group, Universität Hamburg, Mittelweg 177, 20148 Hamburg, Germany cedric.moeller@uni-hamburg.de

Abstract. Existing knowledge graphs are incomplete. A lot of unstructured documents are hiding valuable information. But extracting and structuring that information is expensive. To help, the knowledge graphs can be populated (semi-) automatically. But knowledge graph population methods often assumes existing entities, yet, this is not the reality. To solve this, missing entities need to be detected and distinguished. To support an incoming stream of documents the out-of-KG entities are incrementally modelled. The first goal of the thesis is hence to create a novel entity linking method able to detect, distinguish and incrementally model out-of-KG entities.
While the identification and modelling of potential out-of-KG entities are a step in the right direction, they still need to be included in the knowledge graph. To simplify the process, another goal is to generate short descriptions of newly identified entities. To accomplish that, a method building upon the representation of out-of-KG entities will be created which combines the properties of both graph-to-text and abstractive summarization methods.
For training and evaluation, two silver-standard datasets, as well as one gold-standard dataset, will be created.

Keywords: Knowledge Graph Population · NIL Clustering · Emerging Entity Discovery · Entity Description.

1 Introduction

Today, we are confronted with an ever-increasing number of textual documents in unstructured form. For example, in 2013, every day, around 500 million tweets were published [22]. Now, the number is certainly higher and is further accompanied by other social media posts, news articles and older but now digitized documents. All these documents often contain valuable information but reading them all is no option. Manually extracting relevant information from thousands of documents regarding one's use case is extremely labor-some. To store structured information, KGs are employed in numerous different domains [19]. Having information available in a KG enables the use of powerful services like Question Answering, Recommender Systems, or Reasoning. For instance, the Google search engine relies on an underlying KG [50]. Yet, getting the information, e.g.

from tweets, into the KG is often complicated. Even if an ontology for the KG already exists, domain experts are commonly employed to add new information. Novel knowledge graph population (KGP) methods, also known as knowledge base population, allow that this process can be (semi-) automated [21]. However, existing KGP methods still have several shortcomings which need to be addressed. We will investigate one such shortcoming, which is described next, in the thesis, resulting in two major goals.

Commonly, KGP methods assume that extractable entities and relations are known and part of the underlying KG [2, 8, 26, 53, 56, 60]. But tweets or historical documents often contain entities, which do not exist in a KG yet (**out-of-KG entities**) [29]. Hence, the **recognition** of out-of-KG entities is necessary. To include the entities in the KG in the future, they also need to be put into relation to existing entities. Furthermore, out-of-KG entity might be mentioned repeatedly across documents. To handle this, representations of out-of-KG entities need to be **incrementally modelled**. The **first goal** is therefore the creation of a novel

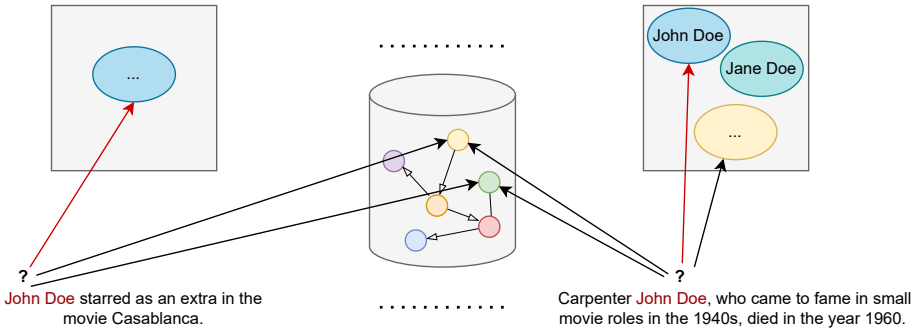

Fig. 1. Out-of-KG entity detection and modelling. An entity mention "John Doe" does not occur in the KG and is identified as an out-of-KG entity, a new intermediary representation is created, which is later used for further linking.

entity linking method supporting out-of-KG entities. It will be able to detect, distinguish and incrementally model out-of-KG entities (see Figure 1). This was, to the best of our knowledge, not pursued to its full potential in research until now.

This brings us to our **second goal**. With accomplishment of the first goal the out-of-KG entities are now identified. But before including them in the KG, it might be wise to check whether they are really suitable. To reduce the effort of this process, another goal of this thesis is the **creation of textual entity descriptions** out of the representations of the new entities. This makes it possible for non-KG experts to quickly understand what a newly identified entity is about, and whether or where to insert it. Furthermore, the description can also be included in the KG itself (see Figure 2). To accomplish that, structured

information in the KG and information in the text needs to be combined. More concretely, a novel method will be developed that will build on existing graph-to-text and text summary methods.

When and how exactly the entities are included in the KG, is out of scope of this work. The focus solely lies on the identification of such entities and creating descriptions. How they are used, will be the responsibility of the curator of the KG.

2 State of the Art

2.1 Knowledge Graph Population

Knowledge Base Population was first defined in TAC-2009 [21, 28] and is concerned with extracting structured information out of text while being constrained to an existing knowledge base. Notably, in TAC-2009 it is also specified that out-of-KG entities should be identified too. However, such entities are not used further. Here, the term Knowledge Graph Population is used as a synonym for KBP. A related task is Open Information Extraction. Here, the goal is to extract structured information while only relying on the input document and no KG. Its main purpose is to pre-process text to improve downstream tasks such as KGP. However, the methods still struggle with coreference resolution and the canonicalization of relations, which are vital for KGP [36].

Most Knowledge Graph Population methods employ a pipeline-based approach [2, 8, 47, 60] using separate entity linking and relation extraction methods. However, as a pipeline always suffers from error propagation, end-to-end approaches are becoming more popular [23, 26, 53, 56]. Not many of those methods consider the possibility of out-of-KG entities. Notable exceptions are KBPearl [23] and the baseline methods used to evaluate the KnowledgeNet dataset [47]. However, they are restricted to only output non-linked triples (by using the entity mentions of the subject and/or object) if an out-of-KG entity is detected. Thus, there is still space for improvement which this thesis desires to fill.

2.2 Out-of-KG Entity Discovery and Representation

Out-of-KG entities were of interest the first time in 2011 when the NIL-clustering task was included in TAC-2011 [21]. NIL-entities are a different term for out-of-KG entities. Here, the goal is to assign the same out-of-KG entities into the same clusters. Important is that all documents are available from the start. Hence, most methods ran clustering algorithms over all documents while calculating different similarity measures between entity mentions [4, 7, 11, 12, 15, 16, 20, 31, 54]. Clustering is usually done by employing rule-based, graph-based or agglomerative-hierarchical clustering methods. The most recent method solving this problem is from 2021 and employs hierarchical clustering over a created mention/entity graph [1]. While offline clustering works well it fails if documents

are not available all at once. This is the case when one has a continuous stream of news or tweets. But also in the context of historical documents, it is the case that not the full batch of documents is available from the get-go. Digitizing and preprocessing documents is a time-consuming endeavor. However, time is precious and waiting until all documents are available can be a critical disadvantage. That is why it is necessary to be able to process the available documents when they arrive.

Hoffart et al. [18] introduced a similar task. However, the focus lies not on out-of-KG entities in general but only on emerging entities. These are recent entities that occur currently in the media and thus are of importance to the public interest. This is typically solved by including auxiliary sources in temporal proximity [18, 59, 62]. Most work focusing on this task is not able to identify the same emerging entities in multiple documents [59, 62]. They are limited to just identifying them. The notable exception is the aforementioned work by Hoffart et al. They represent each emerging entity by the key phrases surrounding it. This makes it possible to link, if the entity linker relies on key phrases, to emerging entities in new incoming documents [18].

This thesis differs in two main aspects. First, the representation will not rely on key phrases but will be based on more informative **dense-embeddings**, second as the support of not only emerging entities but also other out-of-KG entities is of importance, the **availability of auxiliary sources is not assumed**. The feasibility of representing out-of-KG entities by dense embeddings is supported by the impressive performance of recent inductive or zero-shot entity linker relying on such dense embeddings [5, 42, 58]. As no auxiliary sources are assumed to be available, all information used needs to lie in the input documents and KG. Another point differentiating the work of this thesis from all the previous methods is that it will focus on knowledge graphs like Wikidata. Previous methods were only suitable for encyclopedias like Wikipedia or Fandom. This makes different methods necessary and possible.

While similar, the work differs from (cross-document) coreference resolution due to the existence of the large number of entities in the KG which also need to be considered.

Note that there is a debate on what is defined as an entity being desirable to link. Often, this depends on the desired use-case and underlying knowledge graph [43, 44]. As we are concerned with entities not yet existing in the KG, this certainly also affects this work. However, we assume that entity mentions are already available and that any such entity mention does indeed point to an entity of interest, whether in the KG or not. Furthermore, as described in Section 2.3, another goal of the thesis, is the creation of small descriptions of entities. These help to decide whether to include the entity when the entity recognizer is unreliable in identifying entities of interest.

2.3 Entity Description

As the representations of out-of-KG entities will be a combination of structured and unstructured information, the methods to produce a description of the entity

will also need to include the combined properties of methods employed on two common tasks: **Text summarization** and **graph-to-text** generation.

In the past, to tackle graph-to-text generation, sequence-to-sequence models were often employed to map a graph to a text [32, 41, 55]. However, as graphs are unordered by nature, the latest methods usually employ graph neural networks (GNN) to encode the graph and then map the graph representation to output text [3, 10, 27, 33, 39, 40, 51, 64]. This summarizes the state of the art to transform structured information into text.

To transform unstructured text into summaries, two different kinds of text summarization problems are considered: abstractive and extractive summarization. The difference is that extractive summarization chooses certain phrases out of the input document as a summary while abstractive summarization creates an entirely new summary.

As abstractive summarization is more relevant to our problem, the related works in text summarization are dedicated it. Abstractive summarization is a sequence-to-sequence problem which is why most methods follow the encoder-decoder framework [9, 13, 34, 35, 37, 46, 49]. Many different model types, such as pointer networks, convolutional networks or attention-based networks, were employed. In recent years, the best-performing models are based on pre-trained transformer models [6, 25, 48, 61]. Important work is here the one by See et al. [49] as it focuses on the problem of summarization methods often responding with wrong facts. As the planned method of the thesis focuses on summarizations supporting graph information (including factual knowledge) and textual information, producing factual correct summarizes is the goal.

3 Problem Statements, Research Questions and Contributions

There are two problems to solve. First, out-of-KG entities need to be detected and incrementally modelled (see Figure 1). Second, summaries need to be created based on the intermediary representations which are the product of the solution to the first problem (see Figure 2).

Problem 1. Assign each entity mention $m \in M_d$ occurring in an input document d to an entity in one of the three sets $E_{\text{KG}}, E_{\text{I}}, E_{\text{out-of-KG}}$. E_{KG} contains all entities in the KG, E_{I} all entities with intermediary entity representations and $E_{\text{out-of-KG}}$ all other entities. If an entity belonging to the set $E_{\text{out-of-KG}}$ is encountered, create an intermediary representation of such an entity and insert it into E_{I}. It holds that $E_{\text{KG}} \cap E_{\text{I}} = E_{\text{I}} \cap E_{\text{out-of-KG}} = E_{\text{KG}} \cap E_{\text{out-of-KG}} = \emptyset$.

Problem 2. Create a textual summary s of an entity e using the context information of the entity, available in a document d and in the KG. The information in the document is of primary importance in the summary while the KG information provides the holistic frame.

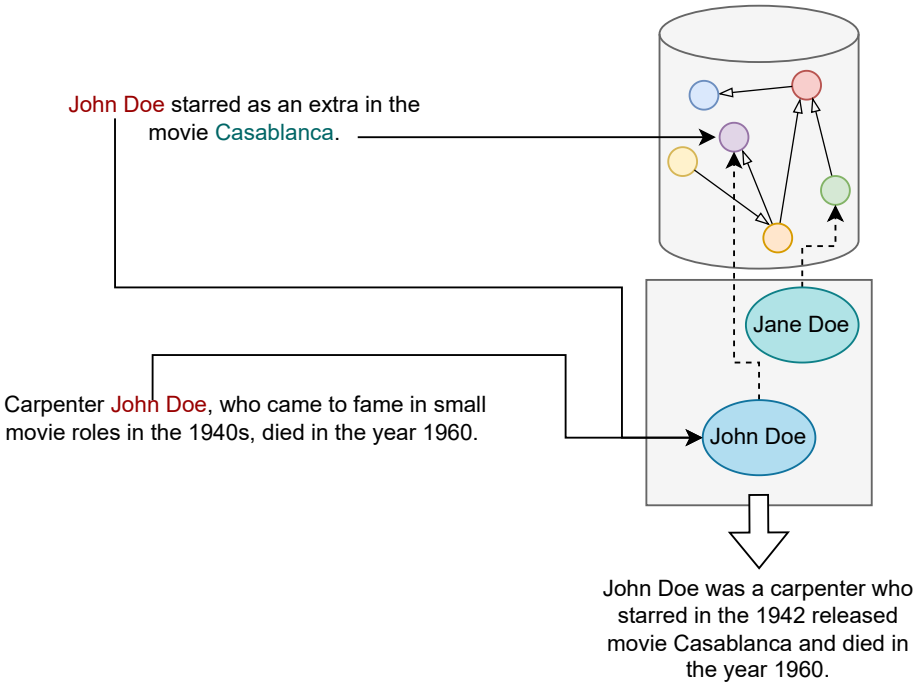

Fig. 2. Out-of-KG entity description creation. Given the intermediary representation of "John Doe", a description is created using the representation and additional information in the KG.

In accordance with the previous two problems the following research questions were identified:
RQ 1: To what extent are out-of-KG entities incrementally identifiable and modellable?
RQ 2: How do existing non-incremental NIL-clustering methods compare to an incremental method?
RQ 3: Does an incremental method utilizing a KG deliver comparable or better performance than emerging entity discovery methods using external documents and encyclopedias?
RQ 4: To what extent are graph-to-text and text summary methods able to be combined to describe out-of-KG entity representations?

In connection with answering the research questions, multiple contributions will be made (more information in Section 5):

- Entity linker supporting out-of-KG entities
- Silver-standard entity linking dataset with focus on out-of-KG entities
- Gold-standard entity linking dataset with focus on out-of-KG entities
- Silver-standard entity description generation dataset
- Gold-standard entity description generation dataset

4 Research Methodology and Approach

The first step is to decide on a suitable dataset for measuring the performance of an entity linker supporting the detection and modelling of out-of-KG entities. While some datasets exist, they are often too small to be able to train complex models on them. That is why a new artificially created dataset is necessary. Hence, potential documents need to be searched for and then processed and labeled to be usable.

To be able to deliver an answer to **RQ 1**, the following steps are followed to design an entity linker supporting out-of-KG entities. Afterward, based on extensive literature review, an entity linking architecture will be created fulfilling both the needs for satisfying entity linking performance and the ability to detect and model out-of-KG entities. The model will be trained and evaluated on the artificially created dataset but also evaluated on the few existing datasets compatible with the problem definition. The entity linking method will be based on one mention encoding model and one entity encoding model. The former model is based on a language model with the input being the document in which each mention is marked via special tokens. This results in a dense embedding of each entity mention. The entity encoding is the concatenation of a label embedding, a type embedding and the output of a graph neural network run over the word embeddings of the n-hop neighborhood of each entity. By not relying on pre-computed graph embeddings, it is on the one hand possible for the entity linking method to link to entities not seen during training and on the other hand it allows to represent out-of-KG entities similar to the in-KG entities. Concerning the out-KG-entity, there are currently two options for the detection and two options for the modelling under discussion. For the detection, the first option is to calculate the ranking score for each potential entity and if the score is below some hyper-parameter, the entity mention is deemed to be out-of-KG. The other option is to use a reinforcement-learning-based approach where the action space consists of all possible entity candidates plus the possibility of an out-of-KG entity. The modelling aspect works as follows: An entity mention is detected to belong to an out-of-KG entity. Then, an artificial graph is built for the out-of-KG entity mention. It consists of the mention itself and all other entity mentions in the document. The entity mentions are linked by extracted relations. Furthermore, for each other entity mention in the document not belonging to an out-of-KG entity, the neighborhood is also added to the artificial graph. Hence, it contains the context information of the document and the graph information of the other entities in the document. One problem is the addition of new information into existing representations as the extension of the corresponding graphs is not straightforward. Also, it might be that an existing representation stands for two entities instead of one, so it is necessary to split the graph which is not trivial. To solve that, the second option will use latent representations in the form of elliptical embeddings, more specifically Gaussian embeddings [17, 57]. When each entity in the knowledge graph and outside of the knowledge graph is embedded as such, incremental clustering methods can

be applied [38, 52]. They enable to seamlessly include new information by adding new clusters, updating them with new samples and even splitting them.

Relation extraction is a smaller focus point and as such a state of the art relation extractor [30, 45, 63] will be used. However, it will be inspected whether the resulting relations can be used to improve the representations of out-of-KG entities. In the case of the symbolic representations in the form of graphs, this is straightforward as the edges between mentions can be refined using the extracted relations. In the case of the latent representations, the inclusion is still under investigation.

After the creation and evaluation of the entity linking method, the representation of out-of-KG entities is specified. Hence, the work on the development of an entity description method can begin. To accomplish that, another comprehensive review of existing abstractive summarization and graph-to-text methods will be carried out. Based on that, a novel method will be designed. As this depends on the literature review, no design decisions are made yet. Here, a dataset does, to the best of my knowledge, not exist. Hence, another dataset will be automatically created from Wikipedia similar to WikiSum [24]. The difference is that here not only documents are assumed as input but also Wikidata as an additional knowledge source. As there exists no comparable work, the evaluation will be done in comparison to created baselines.

Lastly, to evaluate both methods together on a gold-standard dataset, a new one based on historical documents will be annotated. Currently, letters out of the 19th century are considered to be annotated. These will certainly contain out-of-KG entities. However, the letters are not yet available for inspection. If they do not contain enough information on existing entities in, for instance, Wikidata, they are not suitable. The annotation for the EL task will be straightforward. However, for the entity summary task, characteristics defining a "good" summary need to be defined, especially the difference in relevance of the information in the KG and the documents. The plan is to rely on a crowdsourcing platform such as Amazon Mechanical Turk.

See Figure 3 for an overview of the created modules.

5 Evaluation Plan

Entity linking supporting out-of-KG entity detection and modelling is evaluated as follows. As no fully comparable method exists until now, several baselines varying in complexity will be constructed. For example, a simple baseline would be that any entity mention not corresponding to any label of an existing entity, corresponds to an out-of-KG entity, and all such entity mentions correspond to the same entity. A more complex baseline might use the same EL method as the newly designed method but again rely only on entity mention matching to connect out-of-KG entities. However, no final decisions are yet made. The evaluation will be done by using **precision**, **recall** and **F1**. This evaluation will also contribute to answering **RQ 1**. Next, one comparison to NIL-clustering methods and a second to emerging entity discovery methods is planned as these

are related, but not identical, tasks. The former comparison answers **RQ 2**, while the latter will answer **RQ 3**. For NIL-clustering, the commonly used measures, such as **normalized mutual information**, **adjusted rand index** or **B-Cubed+**, will be used. For emerging entity discovery, precision, recall and F1, are again the preferred measures. The method is compared to the baselines on AIDA-EE [18], an automatically created silver-standard dataset and a manually annotated gold-standard dataset. AIDA-EE is also the dataset used for the comparison to the emerging entity discovery methods. For NIL clustering, the methods will be evaluated on the silver-standard and the gold-standard dataset as well as datasets used in NIL-clustering [14].

Lastly, to answer **RQ 4**, the entity description method is evaluated. Here, the **Rouge** score is used, which measures the quality of textual summaries. Baselines need to be created as no comparable methods exist. The baselines will, for example, either only use text or only use graph information. The evaluation will be done on another automatically created silver-standard dataset and the gold-standard dataset based on the same data as used for the entity linking benchmarks.

See Figure 3 for all created datasets and the corresponding modules.

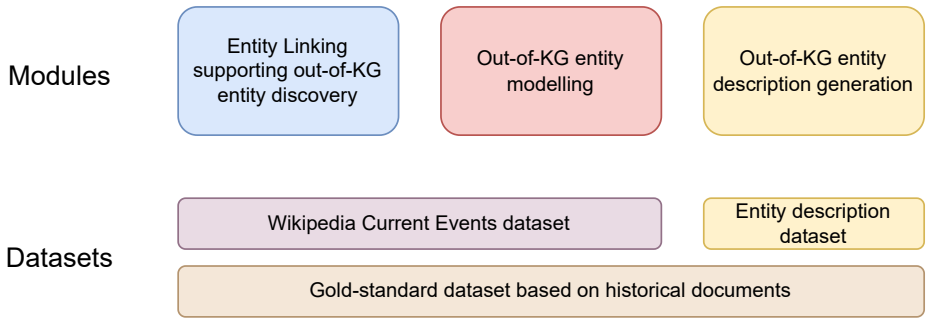

Fig. 3. Overview of created modules and datasets.

6 Preliminary Results

It was already identified that many datasets are not suitable for entity linking with out-of-KG entity detection and the subsequent modelling problem. Often they do not contain out-of-KG entities at all. Other datasets do contain out-of-KG entity mentions but they are only marked as such. It is not specified whether two out-of-KG entity mentions refer to the same entity.

The silver-standard dataset was based on the "Current Events" section of Wikipedia [1]. More specifically, all current events posts between 2017-12-01 and

[1] `https://en.wikipedia.org/wiki/Portal:Current_events`

2022-01-01 were crawled. Then, each hyperlink in the posts was taken as an initial entity mention. To only focus on named entities, each entity mention was further examined. If it is a class, the entity mention was removed. If some document did not contain any entity mention anymore, it was removed. Then, the Wikidata dump of 2017-12-01 was taken and each entity **not** existing in the dump was marked as being out-of-KG. This resulted in a dataset of documents, being sequential in time and containing out-of-KG entities. Statistics on the dataset can be found in Table 1. It is planned to release the dataset together with the first work using the dataset to be confident of its quality.

# examples	23.046
# mentions	64.317
# out-of-KG mentions	7.739
# unique entities	16.175
# unique out-of-KG entities	2.356
Average of # mentions per example	2.8

Table 1. Statistics of Current-Events dataset

7 Conclusion and Future Work

It was argued that while some works on knowledge graph population with out-of-KG entities exist, the topic is still less widespread. Especially regarding the incremental detection and modeling of out-of-KG entities, the research landscape is still rather sparse. Most of the existing work focuses on emerging entities and not out-of-KG entities in general. Furthermore, nearly all such methods have encyclopedias as their target knowledge base.

Hence, one of the main goals will be the creation of a knowledge graph population method with a focus on the incremental detection and modelling of out-of-KG entities. The target knowledge graph will be Wikidata or other graphs which differentiates this work from others focusing on the discovery of out-of-KG entities.

As discovered out-of-KG entities still need to be added to the KG, another goal is the creation of a description of the new entity. While methods exist which can transform graphs to textual descriptions or create summaries of short texts, the combination of both is usually not done. But as out-of-KG entities in our use case are partially grounded in a KG but also part of the textual information of the input documents, combining both subtasks is a necessity and hence another goal.

Acknowledgements

This work is supervised by Prof. Dr. Ricardo Usbeck.

References

[1] Dhruv Agarwal et al. "Entity Linking and Discovery via Arborescence-based Supervised Clustering". In: *CoRR* abs/2109.01242 (2021). arXiv: 2109.01242. URL: https://arxiv.org/abs/2109.01242.

[2] Gabor Angeli et al. "Bootstrapped Self Training for Knowledge Base Population". In: *Proceedings of the 2015 Text Analysis Conference, TAC 2015, Gaithersburg, Maryland, USA, November 16-17, 2015, 2015*. NIST, 2015. URL: https://tac.nist.gov/publications/2015/participant.papers/TAC2015.Stanford.proceedings.pdf.

[3] Daniel Beck, Gholamreza Haffari, and Trevor Cohn. "Graph-to-Sequence Learning using Gated Graph Neural Networks". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 273–283. DOI: 10.18653/v1/P18-1026. URL: https://aclanthology.org/P18-1026/.

[4] Kevin Blissett and Heng Ji. "Cross-lingual NIL Entity Clustering for Low-resource Languages". In: *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*. Minneapolis, USA: Association for Computational Linguistics, June 2019, pp. 20–25. DOI: 10.18653/v1/W19-2804. URL: https://aclanthology.org/W19-2804.

[5] Jan A. Botha, Zifei Shan, and Daniel Gillick. "Entity Linking in 100 Languages". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber et al. Association for Computational Linguistics, 2020, pp. 7833–7845. DOI: 10.18653/v1/2020.emnlp-main.630. URL: https://doi.org/10.18653/v1/2020.emnlp-main.630.

[6] Yue Cao et al. "MultiSumm: Towards a Unified Model for Multi-Lingual Abstractive Summarization". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 11–18. URL: https://aaai.org/ojs/index.php/AAAI/article/view/5328.

[7] Taylor Cassidy et al. "CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description". In: *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*. NIST, 2011. URL: https://tac.nist.gov/publications/2011/participant.papers/CUNY_UIUC_SRI.proceedings.pdf.

[8] Arun Tejasvi Chaganty et al. "Importance sampling for unbiased on-demand evaluation of knowledge base population". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer,

Rebecca Hwa, and Sebastian Riedel. Association for Computational Linguistics, 2017, pp. 1038–1048. DOI: 10.18653/v1/d17-1109.

[9] Sumit Chopra, Michael Auli, and Alexander M. Rush. "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. The Association for Computational Linguistics, 2016, pp. 93–98. DOI: 10.18653/v1/n16-1012.

[10] Marco Damonte and Shay B. Cohen. "Structural Neural Encoders for AMR-to-text Generation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 3649–3658. DOI: 10.18653/v1/n19-1366.

[11] Sourav Dutta and Gerhard Weikum. "C3EL: A Joint Model for Cross-Document Co-Reference Resolution and Entity Linking". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. Ed. by Lluís Màrquez et al. The Association for Computational Linguistics, 2015, pp. 846–856. DOI: 10.18653/v1/d15-1101.

[12] Angela Fahrni et al. "HITS' Monolingual and Cross-lingual Entity Linking System at TAC 2013". In: *Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013*. NIST, 2013. URL: https://tac.nist.gov/publications/2013/participant.papers/HITS.TAC2013.proceedings.pdf.

[13] Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. "Bottom-Up Abstractive Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff et al. Association for Computational Linguistics, 2018, pp. 4098–4109. DOI: 10.18653/v1/d18-1443.

[14] Jeremy Getman et al. "Laying the Groundwork for Knowledge Base Population: Nine Years of Linguistic Resources for TAC KBP". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2018. URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/1047.html.

[15] David Graus et al. "Context-Based Entity Linking - University of Amsterdam at TAC 2012". In: *Proceedings of the Fifth Text Analysis Conference, TAC 2012, Gaithersburg, Maryland, USA, November 5-6, 2012*. NIST, 2012. URL: https://tac.nist.gov/publications/2012/participant.papers/UvA.proceedings.pdf.

[16] Kara Greenfield et al. "A Reverse Approach to Named Entity Extraction and Linking in Microposts". In: *Proceedings of the 6th Workshop on 'Making Sense of Microposts' co-located with the 25th International World Wide Web Conference (WWW 2016), Montréal, Canada, April 11, 2016*. Ed. by Aba-Sah Dadzie et al. Vol. 1691. CEUR Workshop Proceedings. CEUR-WS.org, 2016, pp. 67–69. URL: http://ceur-ws.org/Vol-1691/paper_11.pdf.

[17] Shizhu He et al. "Learning to Represent Knowledge Graphs with Gaussian Embedding". In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*. Ed. by James Bailey et al. ACM, 2015, pp. 623–632. DOI: 10.1145/2806416.2806502.

[18] Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. "Discovering emerging entities with ambiguous names". In: *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*. Ed. by Chin-Wan Chung et al. ACM, 2014, pp. 385–396. DOI: 10.1145/2566486.2568003.

[19] Aidan Hogan et al. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021. DOI: 10.2200/S01125ED1V01Y202109DSK022. URL: https://doi.org/10.2200/S01125ED1V01Y202109DSK022.

[20] Huy M. Huynh, Trong T. Nguyen, and Tru Hoang Cao. "Using coreference and surrounding contexts for entity linking". In: *2013 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future, RIVF 2013, Hanoi, Vietnam, November 10-13, 2013*. IEEE, 2013, pp. 1–5. DOI: 10.1109/RIVF.2013.6719856.

[21] Heng Ji and Ralph Grishman. "Knowledge Base Population: Successful Approaches and Challenges". In: *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. The Association for Computer Linguistics, 2011, pp. 1148–1158. URL: https://aclanthology.org/P11-1115/.

[22] Raffi Krikorian. *New Tweets per Second Record, and How!* https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how. Accessed 2022-03-29. Apr. 2013.

[23] Xueling Lin et al. "KBPearl: A Knowledge Base Population System Supported by Joint Entity and Relation Linking". In: *Proc. VLDB Endow.* 13.7 (2020), pp. 1035–1049. DOI: 10.14778/3384345.3384352. URL: http://www.vldb.org/pvldb/vol13/p1035-lin.pdf.

[24] Peter J. Liu et al. "Generating Wikipedia by Summarizing Long Sequences". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Pro-*

ceedings. OpenReview.net, 2018. URL: https://openreview.net/forum?id=Hyg0vbWC-.

[25] Yang Liu and Mirella Lapata. "Text Summarization with Pretrained Encoders". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, 2019, pp. 3728–3738. DOI: 10.18653/v1/D19-1387.

[26] Yue Liu et al. "Seq2RDF: An End-to-end Application for Deriving Triples from Natural Language Text". In: *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018*. Ed. by Marieke van Erp et al. Vol. 2180. CEUR Workshop Proceedings. CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2180/paper-37.pdf.

[27] Diego Marcheggiani and Laura Perez-Beltrachini. "Deep Graph Convolutional Encoders for Structured Data to Text Generation". In: *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*. Ed. by Emiel Krahmer, Albert Gatt, and Martijn Goudbeek. Association for Computational Linguistics, 2018, pp. 1–9. DOI: 10.18653/v1/w18-6501.

[28] Paul McNamee and Hoa Trang Dang. "Overview of the TAC 2009 knowledge base population track". In: *Text analysis conference (TAC)*. Vol. 17. 2009, pp. 111–113.

[29] Sina Menzel et al. "Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten". In: *Qualität in der Inhaltserschließung*. Ed. by Michael Franke-Maier et al. De Gruyter, Sept. 2021, pp. 229–258. ISBN: 978-3-11-069159-7. DOI: 10.1515/9783110691597-012;https://web.archive.org/web/20220121094046/https://www.degruyter.com/document/doi/10.1515/9783110691597-012/html.

[30] Nandana Mihindukulasooriya et al. "Leveraging Semantic Parsing for Relation Linking over Knowledge Bases". In: *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part I*. Ed. by Jeff Z. Pan et al. Vol. 12506. Lecture Notes in Computer Science. Springer, 2020, pp. 402–419. DOI: 10.1007/978-3-030-62419-4_23. URL: https://doi.org/10.1007/978-3-030-62419-4_23.

[31] Sean Monahan et al. "Cross-Lingual Cross-Document Coreference with Entity Linking". In: *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*. NIST, 2011. URL: https://tac.nist.gov/publications/2011/participant.papers/lcc.proceedings.pdf.

[32] Amit Moryossef, Yoav Goldberg, and Ido Dagan. "Step-by-Step: Separating Planning from Realization in Neural Data-to-Text Generation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 2267–2277. DOI: 10.18653/v1/n19-1236.

[33] Diego Moussallem et al. "NABU - Multilingual Graph-Based Neural RDF Verbalizer". In: *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part I*. Ed. by Jeff Z. Pan et al. Vol. 12506. Lecture Notes in Computer Science. Springer, 2020, pp. 420–437. DOI: 10.1007/978-3-030-62419-4_24. URL: https://doi.org/10.1007/978-3-030-62419-4_24.

[34] Ramesh Nallapati et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. Ed. by Yoav Goldberg and Stefan Riezler. ACL, 2016, pp. 280–290. DOI: 10.18653/v1/k16-1028.

[35] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff et al. Association for Computational Linguistics, 2018, pp. 1797–1807. DOI: 10.18653/v1/d18-1206.

[36] Christina Niklaus et al. "A Survey on Open Information Extraction". In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Association for Computational Linguistics, 2018, pp. 3866–3878. URL: https://aclanthology.org/C18-1326/.

[37] Romain Paulus, Caiming Xiong, and Richard Socher. "A Deep Reinforced Model for Abstractive Summarization". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=HkAClQgA-.

[38] Rafael C. Pinto and Paulo Martins Engel. "A Fast Incremental Gaussian Mixture Model". In: *CoRR* abs/1506.04422 (2015). arXiv: 1506.04422. URL: http://arxiv.org/abs/1506.04422.

[39] Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. "Enhancing AMR-to-Text Generation with Dual Graph Representations". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November*

3-7, 2019. Ed. by Kentaro Inui et al. Association for Computational Linguistics, 2019, pp. 3181–3192. DOI: 10.18653/v1/D19-1314.

[40] Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. "Structural Adapters in Pretrained Language Models for AMR-to-Text Generation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens et al. Association for Computational Linguistics, 2021, pp. 4269–4282. DOI: 10.18653/v1/2021.emnlp-main.351.

[41] Leonardo F. R. Ribeiro et al. "Smelting Gold and Silver for Improved Multilingual AMR-to-Text Generation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens et al. Association for Computational Linguistics, 2021, pp. 742–750. DOI: 10.18653/v1/2021.emnlp-main.57.

[42] Petar Ristoski, Zhizhong Lin, and Qunzhi Zhou. "KG-ZESHEL: Knowledge Graph-Enhanced Zero-Shot Entity Linking". In: *K-CAP '21: Knowledge Capture Conference, Virtual Event, USA, December 2-3, 2021*. Ed. by Anna Lisa Gentile and Rafael Gonçalves. ACM, 2021, pp. 49–56. DOI: 10.1145/3460210.3493549. URL: https://doi.org/10.1145/3460210.3493549.

[43] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. "Fine-Grained Entity Linking". In: *J. Web Semant.* 65 (2020), p. 100600. DOI: 10.1016/j.websem.2020.100600. URL: https://doi.org/10.1016/j.websem.2020.100600.

[44] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. "What Should Entity Linking link?" In: *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management, Cali, Colombia, May 21-25, 2018*. Ed. by Dan Olteanu and Barbara Poblete. Vol. 2100. CEUR Workshop Proceedings. CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2100/paper10.pdf.

[45] Gaetano Rossiello et al. "Generative Relation Linking for Question Answering over Knowledge Bases". In: *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*. Ed. by Andreas Hotho et al. Vol. 12922. Lecture Notes in Computer Science. Springer, 2021, pp. 321–337. DOI: 10.1007/978-3-030-88361-4_19. URL: https://doi.org/10.1007/978-3-030-88361-4_19.

[46] Alexander M. Rush, Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. Ed. by Lluís Màrquez et al. The Association for Computational Linguistics, 2015, pp. 379–389. DOI: 10.18653/v1/d15-1044.

[47] Filipe de Sá Mesquita et al. "KnowledgeNet: A Benchmark Dataset for Knowledge Base Population". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, 2019, pp. 749–758. DOI: 10.18653/v1/D19-1069.

[48] Itsumi Saito et al. "Length-controllable Abstractive Summarization by Guiding with Summary Prototype". In: *CoRR* abs/2001.07331 (2020). arXiv: 2001.07331. URL: https://arxiv.org/abs/2001.07331.

[49] Abigail See, Peter J. Liu, and Christopher D. Manning. "Get To The Point: Summarization with Pointer-Generator Networks". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Ed. by Regina Barzilay and Min-Yen Kan. Association for Computational Linguistics, 2017, pp. 1073–1083. DOI: 10.18653/v1/P17-1099.

[50] Amit Singhal. *Introducing the Knowledge Graph: Things, Not Strings*. https://blog.google/products/search/introducing-knowledge-graph-things-not/. Accessed 2022-03-29. May 2012.

[51] Linfeng Song et al. "A Graph-to-Sequence Model for AMR-to-Text Generation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 1616–1626. DOI: 10.18653/v1/P18-1150. URL: https://aclanthology.org/P18-1150/.

[52] Mingzhou Song and Hongbin Wang. "Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering". In: *Intelligent Computing: Theory and Applications III*. Vol. 5803. SPIE. 2005, pp. 174–183.

[53] Dianbo Sui et al. "Set Generation Networks for End-to-End Knowledge Base Population". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens et al. Association for Computational Linguistics, 2021, pp. 9650–9660. DOI: 10.18653/v1/2021.emnlp-main.760.

[54] Suzanne Tamang, Zheng Chen, and Heng Ji. "CUNY BLENDER TAC-KBP2012 Entity Linking System and Slot Filling Validation System". In: *Proceedings of the Fifth Text Analysis Conference, TAC 2012, Gaithersburg, Maryland, USA, November 5-6, 2012*. NIST, 2012. URL: https://tac.nist.gov/publications/2012/participant.papers/Blender_CUNY.proceedings.pdf.

[55] Bayu Distiawan Trisedya et al. "GTR-LSTM: A Triple Encoder for Sentence Generation from RDF Data". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Mel-*

bourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 1627–1637. DOI: 10.18653/v1/P18-1151. URL: https://aclanthology.org/P18-1151/.

[56] Bayu Distiawan Trisedya et al. "Neural Relation Extraction for Knowledge Base Enrichment". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, 2019, pp. 229–240. DOI: 10.18653/v1/p19-1023.

[57] Luke Vilnis and Andrew McCallum. "Word Representations via Gaussian Embedding". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6623.

[58] Ledell Wu et al. "Scalable Zero-shot Entity Linking with Dense Entity Retrieval". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber et al. Association for Computational Linguistics, 2020, pp. 6397–6407. DOI: 10.18653/v1/2020.emnlp-main.519. URL: https://doi.org/10.18653/v1/2020.emnlp-main.519.

[59] Zhaohui Wu, Yang Song, and C. Lee Giles. "Exploring Multiple Feature Spaces for Novel Entity Discovery". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, 2016, pp. 3073–3079. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12261.

[60] Ce Zhang et al. "DeepDive: declarative knowledge base construction". In: *Commun. ACM* 60.5 (2017), pp. 93–102. DOI: 10.1145/3060586.

[61] Jingqing Zhang et al. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 11328–11339. URL: http://proceedings.mlr.press/v119/zhang20ae.html.

[62] Lei Zhang et al. "Emerging Entity Discovery Using Web Sources". In: *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding - 4th China Conference, CCKS 2019, Hangzhou, China, August 24-27, 2019, Revised Selected Papers*. Ed. by Xiaoyan Zhu et al. Vol. 1134. Communications in Computer and Information Science. Springer, 2019, pp. 175–184. DOI: 10.1007/978-981-15-1956-7_16. URL: https://doi.org/10.1007/978-981-15-1956-7_16.

[63] Yanan Zhang et al. "Adversarial Training Improved Multi-Path Multi-Scale Relation Detector for Knowledge Base Question Answering". In:

IEEE Access 8 (2020), pp. 63310–63319. DOI: 10.1109/ACCESS.2020.2984393.

[64] Chao Zhao, Marilyn A. Walker, and Snigdha Chaturvedi. "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, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, 2020, pp. 2481–2491. DOI: 10.18653/v1/2020.acl-main.224.