Modern Baselines for SPARQL Semantic Parsing

Debayan Banerjee  
debyan.banerjee@uni-hamburg.de  
Universität Hamburg  
Hamburg, Germany

Pranav Ajit Nair*  
pranavajitnair.cse18@itbhu.ac.in  
Indian Institute of Technology (BHU)  
Varanasi, India

Jivat Neet Kaur*  
t-kaurjivat@microsoft.com  
Microsoft Research  
Bengaluru, India

Ricardo Usbeck  
ricardo.usbecker@uni-hamburg.de  
Universität Hamburg  
Hamburg, Germany

Chris Biemann  
christian.biemann@uni-hamburg.de  
Universität Hamburg  
Hamburg, Germany

ABSTRACT

In this work, we focus on the task of generating SPARQL queries from natural language questions, which can then be executed on Knowledge Graphs (KGs). We assume that gold entity and relations have been provided, and the remaining task is to arrange them in the right order along with SPARQL vocabulary, and input tokens to produce the correct SPARQL query. Pre-trained Language Models (PLMs) have not been explored in depth on this task so far, so we experiment with BART, T5 and PGNs (Pointer Generator Networks) with BERT embeddings, looking for new baselines in the PLM era for this task, on DBpedia and Wikidata KGs. We show that T5 requires special input tokenisation, but produces state of the art performance on LC-QuAD 1.0 and LC-QuAD 2.0 datasets, and outperforms task-specific models from previous works. Moreover, the methods enable semantic parsing for questions where a part of the input needs to be copied to the output query, thus enabling a new paradigm in KG semantic parsing. Code and data used for this work can be found at https://github.com/debayan/sigir2022-sparqlbaselines.

CCS CONCEPTS

• Information systems → Question answering.

KEYWORDS

Knowledge Graph, Question Answering, Semantic Parsing, SPARQL

1 INTRODUCTION

Knowledge Graph Question Answering (KGQA) is the task of finding answers to questions posed in natural language, using triples present in a KG. Typically the following steps are followed in KGQA:

1) Objects of interest in the natural language question (NLQ) are detected and linked to the KG in a step called entity linking (EL). 2) The relation between the objects is discovered and linked to the KG in a step called relation linking (RL). 3) A formal query, usually SPARQL, is formed with the linked entities and relations. The query is executed on the KG to fetch the answer. This step is called Query Building (QB), and is the focus of this paper. We assume that gold entities and relations are made available, and our task is to generate the correct SPARQL query. We experiment with a Pointer Generator Network [29] with special vectorisation, and with more recent pre-trained language models such as T5 [26] and BART [22]. We choose models that are able to copy tokens from the input text to the output SPARQL query, apart from being able to handle normal category of questions. Questions which require copy operations are found in the dataset LC-QuAD 2.0 [9], for example:

Is it true that an Olympic-size swimming pool’s operating temperature is equal to 22.4?

which has the corresponding SPARQL over Wikidata KG [42]:

ASK WHERE

{  
wd:Q2084454 wdt:P5066 ?obj  
filter(?obj = 22.4)  
}

We found no previous work for SPARQL QB that can handle such questions. We also show that on datasets where copying is not required (LC-QuAD 1.0 [39]), the approaches still exhibit a strong performance.

Our contribution is as follows:

• We find that with the correct input tokenisation, T5 outperforms all previous works and achieves state of the art performance on LC-QuAD 1.0 over DBpedia [20], and LC-QuAD 2.0 over Wikidata.

2 RELATED WORK

Diefenbach et al. [7] studied several QA systems and concluded that the QB step is generally intertwined with rest of the modules in a pipeline, and hence not evaluated separately. Later, with the advent of Frankenstein [33] and Qanary [32], which are frameworks that allow modular construction of an end-to-end KGQA pipeline, it
was seen that the QB module is the least researched aspect of the pipeline. Soon after, Singh et al. [34] presented a comprehensive survey of individual components of KGQA pipeline and evaluated Sina [31] and NLIWOD\footnote{https://github.com/semantic-systems/NLIWOD/} as individual QB components. Several recent works attempt to solve complex query building by constructing intermediate query representations, such as staged query graphs [6, 16, 19, 23, 43] based on $\lambda$-calculus or skeleton structures [37]. These systems have no built-in method of performing copy operations from input text to output query. The system based on skeleton structure requires manual annotation of queries to corresponding skeletal structures, apart from the final generated query. Some other solutions rely on templates [1, 8, 35], which generally have the natural limitation of the query only being limited to the templates chosen beforehand [2, 41]. In the work of Ding et al. [8], the templates are discovered from the input training data, however, the implementation of the method of discovery of such templates, structures, and sub-structures in the queries is dataset specific. It is unclear if the same method of discovery can scale to all datasets.

Most systems that work over query graphs and use $\lambda$-calculus start with the assumption that entities are already linked, and the relations must still be found. However, with the availability of joint entity and relation linkers such as EARL [10] and Falcon [28], we can develop semantic parsers which focus only on query building and receive both entities and relations pre-linked. For non-KG semantic parsing, PLMs have been evaluated recently with a focus on compositional generalisation [30]. For KG semantic parsing, the Compositional Freebase Questions (CFQ) [18] dataset that is based on the Freebase KG [3], has reported results with non-pre-trained Transformer-based models, while subsequent works [14] [11] over the same dataset have experimented with PLMs. This line of work may be considered closest to ours, however, their focus is on compositional generalisation, while our focus is on the ability to faithfully copy input tokens to the produced query. CFQ has no questions that require such copy operations. Moreover, since Freebase is no longer an active project, our focus is on DBPedia and Wikidata. By exploring two new datasets and corresponding KGs, we expand the scope of research.

Some recent work on Complex Temporal Question Answering [17] handled questions that require extraction of time stamps from the input question. They use task-specific temporal extraction tools such as SUTime [5] and HeidelTime [36] for this purpose, however in the end, they do not form a SPARQL query, and instead attempt to find the correct answer directly through an entity re-ranking approach. Our efforts are instead on finding such flexible single models which can perform such extraction tasks along with the building of SPARQL queries.

The models we experiment with pose no structural or template-based constraints neither at an intermediate stage nor during decoding, and are able to produce any possible query structure since we generate the final query token by token. They require the input question, the linked entities and relations, and a fixed vocabulary of all possible SPARQL tokens. This allows them to operate on all possible datasets and KGs.

3 MODELS

3.1 T5

T5, or Text-to-Text Transfer Transformer, is a Transformer [40] based encoder-decoder architecture that uses a text-to-text approach. Every task, including translation, question answering, and classification, is cast as feeding the model text as input and training it to generate some target text. This allows for the use of the same model, loss function, hyper-parameters, etc. across a diverse set of tasks. We experiment with two variants of the T5 model: T5-small with 60 million parameters and T5-base with 220 million parameters. The uncompressed C4 corpus used to pre-train the models is 750 GB in size.

3.1.1 Input Tokenisation. Let the natural language input question be denoted by $Q = [w_1, w_2...w_n]$ where $w$ denotes the words in the question. The linked entities be denoted by $E = [E_1, E_2...E_n]$ and the relations by $R = [R_1, R_2...R_n]$. For each entity and relation we fetch the corresponding label from the respective KG such that $E_{lab}^1$ denotes the label for $E_1$, and similarly for relations. Let the SPARQL vocabulary be denoted by $V$. The input string to the model is formed as follows:

$$w_1w_2...w_n[SEP]E_1E_{lab_1}...E_nE_{lab_n}[SEP]R_1R_{lab_1}...R_nR_{lab_n}$$

where each word is separated from the other word by a space. We denote the entities by their respective KG IRI’s. For DBpedia the entities take the form, e.g. "http://dbpedia.org/resource/Bolley_Madison" while the relations look like "http://dbpedia.org/ontology/spouse". For Wikidata a typical entity looks like wd:Q76 (Barack Obama) while a typical relation looks like wdt:P31 (instance of). The prefixes wd:, wdt: and several others are expanded according to the list available online\footnote{https://www.wikidata.org/wiki/EntitySchema:E49}. We shuffle the order of entities and relations randomly before passing it to the model. When the input string is passed to T5 as it is, the accuracy is close to zero, as the AutoTokenizer splits the URLs as it sees fit, and it later fails to concatenate the fragments properly at the output. To work around this problem for DBpedia, we made use of the 100 sentinel tokens that T5 provides. The sentinel tokens were originally used during pre-training objective to denote masked tokens, but in our case, we make use of them to represent the prefixes, e.g. http://dbpedia.org/ontology/ as a specific sentinel token such as <extra_id_2>. Additionally, we represent each of the items in the SPARQL vocabulary $V$ as a sentinel token with a specific ID for each keyword. We do the same for prefixes in Wikidata for LC-QuAD 2.0.

3.1.2 Output. The output is composed of tokens from $Q$, $V$, $E$ and $R$. In the case of LC-QuAD 1.0, since there are no questions that require copying of tokens from $Q$, the output is composed of tokens from $V$, $E$ and $R$. We would like to point out that since our input contains the entities and relations, the model always has to perform a form of “copying” even to decide where in the output SPARQL query to place the entities and relations. In the case of questions from LC-QuAD 2.0, in some cases this copying mechanism must
Figure 1: PGN-based QB model. At the current time step, the model is decoding the symbol after the single quote character ('). It considers the scores over the vocabulary and the attention weights over the input text to obtain a final probability distribution, from which it makes the prediction of choosing 1960 as the next token.

also decide which input token from $Q$ to copy and to which position in the output. The output for $E$ and $R$ are not produced verbatim, and instead they are split as a special token prefix, and then the entity or relation ID. A post-processing step is required to resolve these sentinel token IDs into the original prefixes and combine them with the adjacent IDs. For example, wdt:P31 is produced in the output as <extra_id_3> P31.

3.2 BART

The Bidirectional and Auto-Regressive Transformer or BART is a Transformer that combines the Bidirectional Encoder with an Autoregressive decoder into one Seq2Seq model. We experiment with the BART-base model consisting of 139 million trainable parameters. This model was pre-trained on 160 GB of data resulting from the combination of four corpuses, namely, Bookcorpus [45], CC-News [24], OpenWebText [13] and Stories [38].

3.2.1 Input Tokenisation. We form the input string for BART in the same manner as discussed in Section 3.1.1 for T5. However there is no concept of sentinel tokens in BART. Hence to handle special tokens, we instead add them to the BartTokenizer using the add_tokens function and appropriately resize the token embedding space using the resize_token_embeddings function.  

3.3 Pointer Generator Network

A PGN is a seq2seq network with an encoder and a decoder block. We experiment exclusively with LSTM-based [15] PGNs. While PGNs can generate output tokens from a given vocabulary via softmax selection, they are also able to copy tokens from the input text to the output by making use of attentions weights of the decoder. A significant difference between T5, BART and PGN is that for PGNs no pre-training on corpus is performed. The model consists of 53 million trainable parameters which is comparable to T5-small’s 60 million parameters size.

PGNs have typically been employed for abstract summarisation tasks but recently Rongali et al. [27] used PGNs to perform SQL semantic parsing. Figure 1 depicts this architecture parsing an example query.

3.3.1 Input Tokenisation. We construct our input vector (Figure 2) by concatenating the question tokens with gold entities and relations. We add a separator token [SEP] between question tokens

\[ \text{https://huggingface.co/docs/transformers/model_doc/bart} \]
and entity candidates, as well as between entity candidates and relation candidates to help the model understand the segments of the input tokens. Each token is represented by a 968-dimensional vector, where the first 768 dimensions are the BERT contextual embeddings of the question token, or the entity label or relation label, as the case may be. The next 200 dimensions are reserved for the KG embeddings. For question tokens, KG embeddings do not apply, so we fill the 200 dimensions with a sequence of 1.0. The [SEP] token is represented by a vector full of -1.0. The linked entities and relations carry the respective KG embeddings in the last 200 dimensions.

3.3.2 KG Embeddings. For both datasets, we make use of TransE embeddings, due to its popularity and ease of availability [4]. For DBpedia, We train TransE embeddings ourselves using PyTorch-BigGraph [21]. For the training, we proceed to 30 epochs and use the cosine dot operator as the comparator function. For Wikidata we made use of readily available TransE embeddings made available by Pytorch-BigGraph on their webpage.

3.3.3 Re-Ranker. We found that the beams produced by the PGN often contained the right query but not in the top position. To improve the ranking of queries further, we fine-tuned a pre-trained BERT-based classifier (distil-bert-base-uncased) on the output of the PGN. We take the top-10 beam outputs of the PGN on the val sets and query the KG. We save the output of all the queries which produce a valid response from the KG. To train the BERT-based classifier, we form the input string by concatenating the question tokens, the SPARQL query, and the KG response. For supervision, we provide binary labels 0 and 1 for right and wrong query. Once trained, we use the logit values of the penultimate layer of the model for re-ranking the queries produced by PGN on the test set.

### Table 1: Results for query generation with gold entities and relations. Best results are in bold.

|         | F1       | F1       |
|---------|----------|----------|
|         | DBpedia 16.04 | Wikidata |
| LC-QuAD 1.0 | 0.24 | -        |
| NLIWOD  | 0.48     | -        |
| SQG [44]| 0.75     | -        |
| CompQA [23]| 0.77 | -        |
| SubQG [8]| 0.85     | -        |
| AQG-net [6]| -     | 0.45     |
| Multi-hop QLG [19]| -   | 0.53     |
| CLC [46]| -        | 0.59     |
| PGN-BERT          | 0.67 | 0.77     |
| PGN-BERT-BERT     | 0.88 | 0.86     |
| BART              | 0.84 | 0.64     |
| T5-Small          | 0.90 | 0.92     |
| T5-Base           | 0.91 | 0.91     |

4 DATASETS

- **LC-QuAD 1.0** contains 5,000 questions that are answerable on DBpedia 2016-04. The dataset originally contains a 4:1 train-test split. However, several competing systems in Table 1 instead use a smaller split of 3,253 questions. We adopt the same split and follow Ding et al. [8] in performing 5-fold cross-validation with 70:10:20 split for train, dev, and test, respectively. There are no questions that require copying of input tokens to output query in this dataset.

- **LC-QuAD 2.0** is based on Wikidata and consists of a mixture of simple and complex questions that were verbalized by human workers on Amazon Mechanical Turk. It is a large and varied dataset comprising 24,180 train questions and 6,046 test questions. In addition, to the best of our knowledge, this is the sole dataset that incorporates aspects of hyper-relational [12] structure of recent Wikidata versions. Approximately 16% of questions in the dataset require copying of input tokens to the output query.

5 EVALUATION

In Table 1, the results of competing systems for LC-QuAD 1.0 are shown as reported by Ding et al. [8]. The results for competing systems for LC-QuAD 2.0 are as reported by Zou et al. [46], also found on the KGQA leaderboard [25]. For evaluating T5, BART and PGNs on both datasets, we take the top-10 beams outputs of the models and query the KG in ranked serial order. The first query to return a non-empty response is considered the output of the model. We match the KG responses of the generated query to the gold query and mark it a match if they are identical, and no match if they are not.

In the case of LC-QuAD 2.0, we found a large number of queries no longer answerable on the current version of Wikidata. The original dump used to create the dataset is no longer available online. As a result, we setup an endpoint with the dump dated 13 October 2021 of Wikidata and filtered the test set of 6046 questions down to 4211 questions for which the gold query produced a valid response. We loaded the triples to a Virtuoso Open Source Triple Store and used it as our endpoint. Zou et al. [46] resort to a similar pre-filtration step for the dataset in their work.

In Table 1, PGN-BERT refers to the setting where we feed the PGN with BERT embeddings. We apply no re-ranker to the output beams, and consider the first query in the ranked beams to produce a valid response from the KG as the output of the system. On the other hand, PGN-BERT-BERT refer to the setting where the beams are further re-ranked with the BERT-based re-ranker described previously.

6 DISCUSSION

T5 not only outperforms all previous works on both datasets, but also BART and PGN that we experiment with. It is important to handle input tokenisation of prefixes through sentinel IDs for T5, or else the model produces zero accuracy. In the case of PGN, we also performed a test without input KG embeddings, where we

---

3 https://kgqa.github.io/leaderboard/
4 https://databus.dbpedia.org/dbpedia/wikidata/debug/2020.07.01
5 https://dumps.wikimedia.org/wikidatawiki/entities/
6 https://github.com/facebookresearch/PyTorch-BigGraph
saw a 20% point drop in accuracy. In spite of the addition of KG embeddings, it fails to get an edge over T5, which operates solely based on entity and relation labels. It can hence be concluded that the large amount of pre-training for T5 outperforms the KG embedding advantage.

In the case of BART versus T5, we observed that the copy actions of BART were not as clean as T5. For example, in LC-QuAD 2.0, BART produced an F1 of 0.3 for questions that require token copy to the output, while T5 produced an F1 of 0.8. We believe this is due to the fact that instead of using sentinel IDs, which are inherently a part of the T5 model from its pre-training objective, we had to resort to adding special tokens to the BartTokenizer, which does not produce the same copying performance.

T5-small and T5-base produce similar performance, which suggests that the extra number of parameters in the base model remains unused for our given task. For LC-QuAD 2.0, none of the previous works has the ability to copy tokens from input to the output, so our approaches outperform them by far. In terms of PGN performance, on using the BERT re-ranker, we see gains in the range of 10-20 % points, which shows that PGN does a poor job of producing the right query towards the top. When comparing T5-small1 performance with PGN, although they have similar parameter count, the pre-training on 750 GB of corpus seems to be the crucial factor in T5-small’s favour.

### ERROR ANALYSIS

We randomly sampled 100 cases of erroneous outputs for PGN, T5-Small and BART, as shown in Table 2. The most important difference in the errors produced by T5/BART versus PGN is what we call "copy morphing". T5/BART produces their output from an open ended sub-word-based vocabulary, and hence sometimes corrupt the item being copied. At times, they produce wrong entity IDs, e.g., in the case of DBpedia, Barack-Obama instead of Barack_Obama, and in some other cases, they hallucinate entity IDs that are not part of the linked entities or relations in input. We also encountered cases of morphing relation IDs to semantically similar words, like notableWork changing to notabilityWork, or unexpected capitalisation, e.g., Artist instead of artist. Although PGN’s overall accuracy is lower, they do not exhibit this issue, since the copying takes place from an expanded vocabulary which is limited, consisting of SPARQL tokens and the input tokens. We also found that copy morphing in T5 takes place only for DBpedia where the entity and relation IDs are similar to the dictionary labels. For Wikidata, where the IDs are numerical, morphing does not occur.

We categorise the other kind of errors under the following heads: Triple Flip, e.g., <s p s> instead of <s p o>, Wrong Variables, e.g., ?varO instead of ?var1, Wrong Intent, e.g., ASK instead of SELECT, Copy Errors, where the wrong token is copied, and Syntax Errors, where the generated query is malformed and an invalid SPARQL query. We observe that BART has poor performance in copying, while other models have most errors in triple flips and wrong variables.

### CONCLUSION AND FUTURE WORK

In this work, we establish new baselines using PLMs for the KG semantic parsing task for two datasets. Our results surpass the results of earlier baselines and lay grounds for future research. We make use of readily available and popular models, with minimal changes to input vectorisation and tokenisation, to report state of the art results on SPARQL semantic parsing and query building tasks over DBpedia and Wikidata. We show that these methods are flexible and able to handle a variety of questions with no fixation on templates or query graphs. The direction of future semantic parsing work for SPARQL should lay more emphasis on pre-training, since currently it is more focused on producing custom model architectures for the tasks.

As future work, we would like to explore the ability of these models in disambiguation tasks, where the input consists of entity and relation candidates, instead of linked entities and relations.

### ACKNOWLEDGMENTS

This research was partially funded by the German Federal Ministry of Education and Research (BMBF) as part of the INSTANT project, ID 02L18A111.

### REFERENCES

[1] Abdalghani Abujabal, Mohamed Yahya, Mirek Riedewald, and Gerhard Weikum. 2017. Automated Template Generation for Question Answering over Knowledge Graphs. International World Wide Web Conferences Steering Committee.

[2] Ram G. Athreya, Srividya Kona Bansal, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. 2021. Template-based Question Answering using Recursive Neural Networks. In 15th IEEE International Conference on Semantic Computing, ISCSC 2021, Laguna Hills, CA, USA, January 27-29, 2021. IEEE.

[3] Kurt Bolbicker, Colin Evans, Praveen Paritosh, Tim Starge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery. https://doi.org/10.1145/1376616.1376746

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakunenko. 2013. Translating Embeddings for Modeling Multi-Relational Data (NIPS). Curran Associates Inc.

[5] Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12). European Language Resources Association (ELRA).

[6] Yongrui Chen, Huiling Li, Yuncheng Hua, and Guilin Qi. 2020. Formal Query Building with Query Structure Prediction for Complex Question Answering over Knowledge Base. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization.

[7] Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Mar et. 2018. Core Techniques of Question Answering Systems over Knowledge Bases: A Survey. Knowledge and Information Systems (2018).

[8] Jiwei Ding, Wei Hu, Qixun Xu, and Yuhong Qu. 2019. Leveraging Frequent Query Substructures to Generate Formal Queries for Complex Question Answering. 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). ACL.

[9] Mohinsh Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over
