Towards Formally Grounded Evaluation Measures for Semantic Parsing-based Knowledge Graph Question Answering

Trond Linjordet
University of Stavanger
Stavanger, Norway
trond.linjordet@uis.no

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Vinay Setty
University of Stavanger
Stavanger, Norway
vinay.j.setty@uis.no

ABSTRACT

Knowledge graph question answering (KGQA) is important to make structured information accessible without formal query language expertise on the part of the users. The semantic parsing (SP) flavor of this task maps a natural language question to a formal query that is machine executable, such as SPARQL. The SP-KGQA task is currently evaluated by adopting measures from other tasks, such as information retrieval and machine translation. However, this adoption typically occurs without fully considering the desired behavior of SP-KGQA systems. To address this, we articulate task-specific desiderata, then develop novel SP-KGQA measures based on a probabilistic framework. We use the desiderata to formulate a set of axioms for SP-KGQA measures and conduct an axiomatic analysis that reveals insufficiencies of established measures previously used to report SP-KGQA performance. We also perform experimental evaluations, using synthetic and state-of-the-art neural machine translation approaches. The results highlight the importance of grounded alternative SP-KGQA measures.

CCS CONCEPTS

• **Information systems** → **Retrieval effectiveness**; **Question answering**.

KEYWORDS

Knowledge graph question answering; evaluation measures; axiomatic IR evaluation

ACM Reference Format:
Trond Linjordet, Krisztian Balog, and Vinay Setty. 2022. Towards Formally Grounded Evaluation Measures for Semantic Parsing-based Knowledge Graph Question Answering. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539813.3545146

1 INTRODUCTION

Question answering systems provide a way for users to express their information needs in a natural manner [46, 48]. By harnessing structured data in the form of knowledge graphs (KGs), knowledge graph question answering (KGQA) can facilitate information access that would otherwise require expertise in formal query languages. The task of KGQA is typically approached as either an *information retrieval* (IR) or a *semantic parsing* (SP) problem [18, 33]. In addition, some hybrid approaches have been attempted [31, 38]. IR-based KGQA has the advantage of being able to robustly leverage entity descriptions as well as graph structure. However, it is difficult to interpret how the system arrived at the retrieved answers, and to verify whether the system's "understanding" of the question was correct. Conversely, SP-based KGQA (SP-KGQA) predicts an explicit formal query (e.g., SPARQL) that represents the natural language question (NL question) posed by a human user, and, in turn, executes the formal query to retrieve answers [18, 33]. This provides greater interpretability by showing explicitly how the system "understood" the NL question. This means the reasoning represented by the formal query can be reconstructed in natural language by a human who is an expert in the formal query language. In other words, the interpretation may lie beyond the ability of non-expert users in the case of complex formal queries, but the fact that a single formal query is expressed allows interpretation in principle. The reported evaluation measures in KGQA research are often repurposed from the evaluation measures of other tasks, like machine translation and ad hoc retrieval [18, 33]. When the interpretability of predictions is of interest, i.e., SP-KGQA, it is not clear that straightforward adoption of measures established for other tasks is appropriate. In fact, SP-KGQA evaluation has been an undeservedly neglected field of research.

In order to capture the performance quality of SP-KGQA systems there is a need for a theoretically grounded analysis which is currently missing in the field. Axiomatic analysis has been a productive methodology to investigate and develop evaluation measures [1–3, 5, 6, 15, 25–28, 50, 51]. In the present work, we therefore make an initial effort to apply the axiomatic approach to developing formally grounded evaluation measures for SP-KGQA.

Research in SP-KGQA is increasingly oriented towards neural machine translation (NMT) architectures [23, 56, 65], and the reported results are promising. We therefore further limit our scope to focus on state-of-the-art NMT methods in our experimental evaluation, noting that the proposed measures are nevertheless applicable to all SP-KGQA systems. The scope of NMT systems is chosen as a tractable and sufficient experimental scope where state-of-the-art performance from previous work with open-source codebases can be reproduced on a complex KGQA dataset.

We begin by describing the desiderata of the SP-KGQA task. Using these desiderata, we derive a probabilistic framework for novel compound measures, a number of specific component measures,

ICTIR '22, July 11–12, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545146

and specific instantiations of the framework, i.e., novel compound measures.

We next postulate axioms for SP-KGQA measures on the basis of the task desiderata. With this theoretical basis, we perform an axiomatic analysis on established measures, as well as novel proposed measures. This analysis reveals that all the established measures used to evaluate SP-KGQA in previous work have critical shortcomings with respect to properly evaluating this task, as established measures can only partially satisfy the axioms.

To validate the framework itself and the novel measures, we perform an experimental evaluation using both synthetic (ground truth degraded in a controlled manner) and state-of-the-art NMT SP-KGQA models. We find that important differences between NMT architectures for SP-KGQA are obscured by evaluating with individual established measures. Our proposed measures provide the necessary instrumentation to conduct a balanced re-evaluation.

The main contributions of this paper are the following:

(1) Development of novel SP-KGQA evaluation framework and novel measures;
(2) First formulation of axioms for SP-KGQA measures;
(3) Axiomatic analysis of established and novel measures;
(4) Extensive empirical evaluation using both synthetic runs and learned models.

The resources accompanying the paper are made available at https://github.com/iai-group/ictir2022-kgqaeval.

2 RELATED WORK

We review work on knowledge graph question answering and axiomatic analysis of evaluation measures.

2.1 Knowledge Graph Question Answering

Question answering (QA) over unstructured text has been the focus of research for decades within the fields of information retrieval (IR) and natural language processing. However, in recent years, QA over structured data sources such as knowledge graphs (KGQA) has gained popularity [20, 48, 60]. One prominent approach to KGQA is *semantic parsing* (SP) [7, 18, 33, 40].[1] Under this approach (SP-KGQA), the task is to transform a natural language (NL) question to a structured formal query, such as SPARQL, which can then be executed against a KG to retrieve the answer.

SP-KGQA has long been studied as a problem with distinct subtasks to be solved in modular manner [33, 36]. However, state-of-the-art SP-KGQA systems have also shown promising results using end-to-end trained neural machine translation (NMT) techniques [23, 56, 65].

There exist several KGQA benchmarks, which can be broadly classified as (a) *simple*, consisting of a single entity mention and a single relation which require only one KG fact for answering, and (b) *complex*, needing additional conditions and more than one KG fact to answer. Examples of simple KGQA benchmarks include WebQuestions [11], SimpleQuestions [13], and Free917 [17] over Freebase, and SimpleQuestions [21] over Wikidata. Recently, the focus has been shifted to complex KGQA benchmarks [44, 58], examples of which include DBNQA [30] over DBpedia, LC-QuAD 2.0 [24] over DBpedia and Wikidata, ComplexWebQuestions [58], ComplexQuestions [9], and GraphQuestions [57] over Freebase. Of these, we choose to focus on DBNQA [30] because it is the largest complex dataset that has been used for evaluating several NMT-based models [65].

KGQA systems have been evaluated using typical IR measures [39], either considering the correctness of answers [10, 36, 53], ranked candidate formal queries generated [22, 38], or in terms of subtasks [54], such as entity linking [64] and answer type prediction [41]. In evaluating answers with typical IR measures, set-based measures (e.g., accuracy, precision, recall, and F_1) [10, 19, 22, 45, 53, 64] and ranked-list-based measures (e.g., H@1, MRR) [16] have been used. We note that evaluating answers using a ranked-list approach may not be suitable, since the user relies on the system to provide a single (definitive) answer as opposed to a list of candidate answers. This may be a consequence of the IR-based approach commonly taken with the task, in order to award partial reward in the evaluation. Instead, we devise ways to measure partial success on a single prediction rather than a list of candidate predictions.

In contrast, some existing SP-KGQA systems and benchmarks also report machine translation-based evaluation measures [56, 65] with respect to the predicted formal query. Those measures, like BLEU [42], focus on n-gram overlap, which is insufficient to capture the complexities of formal queries. Regardless, the appropriateness of these measures has not been addressed to date. This paper aims to fill that gap.

2.2 Deriving Measures Axiomatically

There is a solid body of existing research on axiomatically deriving evaluation measures for various information access tasks. In this approach, formal constraints are defined and used to theoretically show which performance measures satisfy each constraint, and hence possess the corresponding quality. Several tasks have been studied in this way, including clustering [3], classification [50], filtering [4], quantification [51], diversification [1, 6, 49], and recommender systems [43]. In addition, the axiomatic methodology itself has been investigated in the context of IR [2] and the properties of IR effectiveness measures have been axiomatically analyzed [15, 26], such as whether they are interval scales [27] and what statistical properties different measures have as a consequence [25]. Finally, work has been done towards constructing general theories of IR effectiveness measurements [5, 28]. In this paper, we make the first attempt to axiomatically derive evaluation measures for the SP-KGQA systems.

3 MEASUREMENT FRAMEWORK

We begin by defining the SP-KGQA task and associated desiderata in Sect. 3.1. On this basis, we derive a probabilistic framework to express an ideal SP-KGQA measure in Sect. 3.2. We next develop component measures in Sect. 3.3 and present some possible compound measures in Sect. 3.4.

3.1 The Semantic Parsing KGQA Task

The task of SP-KGQA is, given some natural language (NL) question q, to produce a corresponding formal query f that when executed on knowledge graph $\mathcal{K}$ will return the correct answer a to the

[1] In fact, Chakraborty et al. [18] claim this is the most common approach to KGQA.

question. We denote the result of this query execution as $r_{\mathcal{K}}(f) = a$. If the formal query is syntactically incorrect, attempted execution returns an error, $r_{\mathcal{K}}(f) = \epsilon$. Conversely, if the formal query is well-formed but returns no entities or values, we denote this as $r_{\mathcal{K}}(f) = \emptyset$. In the present work we only address NL questions that can be correctly represented as a KG formal query, e.g., factual questions with semantic support in the KG. In the context of a specific $\mathcal{K}$ then, for an NL question q there exists a corresponding ground truth formal query f^* that represents the NL question in a structured manner (i.e., as a logical form), which when executed returns the ground truth answer, $r_{\mathcal{K}}(f^*) = a^*$.

According to Chakraborty et al. [18], a correct formal query f produced by SP-KGQA must both correctly represent the meaning of q, and return the correct answer a corresponding to q when f is executed on the KG. We write $f \overset{\text{sem}}{\equiv} f^*$ to denote semantic equivalence between a predicted formal query f and the ground truth formal query f^* (acknowledging that f may be semantically equivalent to f^* without being verbatim identical). As a simple example, a fact `x parentOf y` is semantically equivalent to the fact `y childOf x`. Another example of semantic equivalence is that the ordering of triple patterns in the `WHERE`-clause of a SPARQL query can be reordered without changing the meaning or effect of the formal query:

```
SELECT ?person WHERE { ?person childOf ?parent
 . ?parent childOf Albert_Einstein }
```

is semantically equivalent to

```
SELECT ?person WHERE {?parent childOf Albert_Einstein
 . ?person childOf ?parent }
```

and both these formal queries represent the NL question "Who are the grandchildren of Albert Einstein?"

Note that retrieving the correct answer requires the formal query to be syntactically correct. We therefore also make the *executability* requirement explicit in postulating the following desiderata:

- D1 *Semantic representation* (or semantic structure in [18]): f correctly represents the meaning of q. Formally: $f \overset{\text{sem}}{\equiv} f^*$.
- D2 *Syntax correctness:* f is well-formed under the formal query language and does not return an error ϵ when executed on $\mathcal{K}$. Formally: $r_{\mathcal{K}}(f) \neq \epsilon$.
- D3 *Answer correctness:* f when executed on $\mathcal{K}$ retrieves the correct answer a. Formally: $r_{\mathcal{K}}(f) = r_{\mathcal{K}}(f^*) = a^*$.

An imperfect system could satisfy some but not all these desiderata. For example, if the NL question is "What country has the highest GDP in the world?", then a good system would retrieve the answer "USA." If a system retrieves the same answer by a formal query that represents the meaning of a different NL question, such as "What country has a flag known as the Stars and Stripes?", this would be better than getting the wrong answer, but not as good as getting the correct answer with the correct formal query.

We want correct answers, but a reliable and trustworthy system must get the correct answers for the correct reasons. Thus, syntax correctness reflects whether the formal query is well-formed, and hence executable, while semantic representation reflects how well the meaning of the NL question is reflected in the formal query.

As a more explicit example from our dataset and experiments, we can compare a ground truth formal query with a spurious predicted formal query that illustrates the same point: The NL question is "Did justin madden study at the australian catholic university university?" [sic] and the ground truth formal query is

```
ASK WHERE { dbr:Justin_Madden dbp:university
 dbr:Australian_Catholic_University }
```

but the model predicts the formal query

```
ASK WHERE { dbr:Justin_Madden dbo:almaMater ?uri . }
```

which represents a different NL question, "Did Justin Madden have an alma mater?" However, the answers to the ground truth and predicted formal queries are identical: `True`. This illustrates how measuring task performance only with respect to one desideratum could give a misleading impression, either overestimating the quality of predictions or neglecting the merits of imperfect performance.

3.2 Probabilistic Framework

Given that the SP-KGQA task has several desiderata which might not all be fulfilled simultaneously, an evaluation measure for SP-KGQA should ideally take each of the three desiderata into account. Furthermore, we would like to be able to quantify partial success with respect to each desideratum. Therefore, the measure of each desideratum may be expressed as a probability of how plausible [32] the prediction is, given the evidence. Here, the prediction comprises both the predicted formal query and answer, while the evidence referenced is the ground truth formal query and answer. We denote the overall correctness of the SP-KGQA prediction as the probabilistic expression $P(C = 1|q, f, f^*)$.

Here C is a binary random variable denoting correctness with respect to all desiderata, $C = D_1 \wedge D_2 \wedge D_3$, where D_i is the binary random variable corresponding to desideratum Di. For binary random variable $X \in \{0, 1\}$, we simplify the probability notation as $P(X = 1) = P(X)$. We then express $P(C|q, f, f^*)$ as:

$$P(C|q, f, f^*) = P(D_1 \wedge D_2 \wedge D_3 \mid q, f, f^*), \tag{1}$$

$$= P(D_1|q, f, f^*)\, P(D_2 \wedge D_3|q, f, f^*) \tag{2}$$

$$= P(D_1|q, f, f^*)\, P(D_2|q, f, f^*)\, P(D_3|D_2, q, f, f^*) \tag{3}$$

$$= P(D_1|f, f^*)\, P\big(D_2|r_{\mathcal{K}}(f)\big)\, P\big(D_3|D_2, r_{\mathcal{K}}(f), r_{\mathcal{K}}(f^*)\big) \tag{4}$$

Starting from Eq. (1) we make the assumption that D1 is independent both from D2 and from D3, and get Eq. (2). This assumption is made because predicted formal queries can represent relevant semantic and structural aspects of the NL question, even if other desiderata are not well satisfied. As stated, if the predicted formal query is not executable, there can be no answers, much less correct answers. Therefore, D3 is conditionally dependent on D2. In contrast, it is perfectly possible for a predicted formal query to have correct syntax without returning the ground truth answer. Therefore D2 is conditionally independent from D3. This yields Eq. (3) from Eq. (2). Finally, in Eq. (4) we express each *component* in terms of the input variables immediately relevant to them.

Equation (4) presents a general framework, based on which one can instantiate specific *compound* measures, by setting the different components. Note that the instantiated compound measure does not need to be strictly probabilistic; since component measures are multiplied, any real-valued measure in a fixed range may be used. The framework is probabilistic to make explicit the dependencies between the desiderata. Nevertheless, the component measures do not need to be probabilistic—this allows us to instantiate specific measures based on the framework using existing measures as component measures (as we do with BLEU in Sect. 3.4), with the overall result being rank-equivalent.

In addition, the framework can be relaxed to ignore a desideratum by substituting the value 1 for the respective component. Also note that because of the multiplicative formulation, the compound expression in Eq. (4) yields 0 if any of the components are 0. In order to preserve the evaluation of partial successes on each component, we need to make sure that $P(D_i|.) > 0$ for $i \in 1, 2, 3$. This could be ensured by enforcing a minimum non-zero value γ for each of the components: $P(D_i|.) = \gamma + (1 - \gamma)\hat{P}(D_i|.)$. Finally, this formulation assumes that each of the desiderata are weighted equally. However, the framework could be extended to allow for non-uniform weighting; this is left to future work.

3.3 Component Measures

Next, we consider in more detail each component and discuss suitable measures, using either established measures, where appropriate, or developing new ones.

Measuring Semantic Representation: We express D1 as $f \overset{\text{sem}}{\equiv} f^*$ to emphasize that semantic equivalence is more nuanced than just an Exact Match [23, 47, 65]. From machine translation, *n*-gram-based measures like BLEU [42] or ROUGE [34] may be used to measure semantic representation, capturing both partial and complete success of a predicted formal query. However, such *n*-gram-based approaches do not distinguish between terms; specifically, they do not recognize the key semantic elements in the formal query, like entities and predicates. We therefore define novel component measures to quantify semantic representation, $P(D_1|q, f, f^*)$, that can reflect both partial and complete success in terms of semantic elements in the formal query. First, considering the formal queries as sets of individual semantic elements, i.e., entities and predicates: $f_{\text{Sem}} = \{e \in f\} \cup \{p \in f\}$.

This gives us recall and precision on the level of individual semantic elements,

$$R_{\text{Sem}}(f, f^*) = \frac{|f_{\text{Sem}} \cap f^*_{\text{Sem}}|}{|f^*_{\text{Sem}}|}; \quad P_{\text{Sem}}(f, f^*) = \frac{|f_{\text{Sem}} \cap f^*_{\text{Sem}}|}{|f_{\text{Sem}}|} \tag{5}$$

and consequently semantic representation F-measure,

$$F_{\beta,\,\text{Sem}}(f, f^*) = \frac{1 + \beta^2 R_{\text{Sem}}(f, f^*) P_{\text{Sem}}(f, f^*)}{R_{\text{Sem}}(f, f^*) + \beta^2 P_{\text{Sem}}(f, f^*)}, \tag{6}$$

where we simply take $\beta = 1$ to have $\text{F}_{1,\,\text{Sem}}(f, f^*)$.

Second, following the same rationale as above, but considering formal queries as sets of (s, p, o) triples patterns: $f_{\text{Tri}} = \{(s, p, o) \in f\}$, we define the triple patterns-based semantic representation F-measure where we simply take $\beta = 1$ to get $\text{F}_{1,\text{Tri}}(f, f^*)$, following the same steps as Eq. (6). Following Usbeck et al. [62], for each set-based measure, if both the ground truth answer and the predicted answer are empty sets, the score for an instance is 1.0. Note that we here take advantage of the permutation invariance of the triple patterns as illustrated in Sect. 3.1. For both $\text{F}_{1,\,\text{Sem}}(f, f^*)$ and $\text{F}_{1,\text{Tri}}(f, f^*)$, the rationale is to focus on the distinct meaningful parts of the formal query, i.e., the URIs, while disregarding the syntactical elements and triple pattern ordering. In the case of $\text{F}_{1,\text{Tri}}(f, f^*)$, the evaluation is simplified by ignoring the placeholder variables in the triple patterns, which would otherwise require more involved coordination reflecting the graph structure of the formal query. Extending $\text{F}_{1,\text{Tri}}(f, f^*)$ in this regard may be warranted in future work.

Measuring Syntax Correctness: Simply and strictly, syntax correctness of the predicted formal query can be evaluated by execution. We distinguish this specific measure *Executability* (Exec), from D2 (*syntax correctness*) as it may be possible to measure degrees of syntax correctness. Hence, we express D2 simply as $r_{\mathcal{K}}(f) \neq \epsilon$. We then have $\text{Exec}(f) = 1$ iff $r_{\mathcal{K}}(f) \neq \epsilon$, otherwise $\text{Exec}(f) = 0$. A continuous measure of syntax correctness with values between 0.0 and 1.0 could be implemented in various ways, but this is left as future work.

Measuring Answer Correctness: For answer correctness, Exact Match is an applicable established measure. However, to capture partial success, we can consider the retrieved answers as sets of result tuples, $\mathcal{T}_a = \{\tau_a \in a\}$. We can then obtain an answer F-measure, following the same steps as Eq. (6), to obtain $\text{F}_{1,\text{Ans}}$. As D3 depends on D2, if $r_{\mathcal{K}}(f) = \epsilon$ we simply set $P(D_2|.) = P(D_3|.) = \gamma$.

There is a variety of types of answers that can be retrieved by formal queries from $\mathcal{K}$. For a given formal query (considering formal queries in SPARQL, specifically), the answer type may be, for example, a boolean, an entity or predicate URI, a literal value, a tuple, or a set of tuples. Generally, we treat all answers as sets of tuples, even if they contain a single item. That way we can use set overlap-based measures for answer correctness.

3.4 Novel Compound Measures

Using the probabilistic evaluation framework introduced in Sect. 3.2, we now instantiate three novel compound measures GEK-1..3, where GEK is an acronym for "Grounded Evaluation of SP-KGQA." Specifically, we vary the semantic representation component, but measure syntax correctness and answer correctness in a fixed way, yielding the following novel compound measures:

$$\begin{aligned}
\text{GEK-1} &= \text{BLEU} \cdot \text{Exec} \cdot \text{F}_{1,\text{Ans}} \\
\text{GEK-2} &= \text{F}_{1,\text{Sem}} \cdot \text{Exec} \cdot \text{F}_{1,\text{Ans}} \\
\text{GEK-3} &= \text{F}_{1,\text{Tri}} \cdot \text{Exec} \cdot \text{F}_{1,\text{Ans}}
\end{aligned}$$

While other combinations would also be possible, we have selected a tractable number of compound measures that are expected to address the shortcomings of established measures.

4 AXIOMATIC ANALYSIS

We have proposed a framework for SP-KGQA measures based on task desiderata in Sect. 3.1. In light of these, we develop axioms that formally express the constraints that SP-KGQA measures should satisfy. We then analyze established and proposed measures in terms of these axioms.

4.1 Axioms

The SP-KGQA task consists of instances with elements (q, f, f^*), where for a given KG (f, f^*) determine the answers (a, a^*). For each axiom, we consider an abstract measure of the form $m(f, f^*)$, the evaluation of a predicted formal query f with respect to the ground truth f^*. The axioms discuss comparisons of the form $m(f_1, f^*) \geq m(f_2, f^*)$ when comparing the evaluation properties of two hypothetical predictions f_1 and f_2. Note that answer-level measures are analogously expressed in the form $m(a, a^*)$.

Our first axiom (A1) corresponds to desideratum D1. Practically, we must evaluate the logical equivalence of f to q by comparing f to f^*.

Axiom A1 - Semantic representation: A formal query f may be considered as a set $\mathcal{U}_f$ of elements u_f that are semantic properties extracted from the formal query f, such as entities in f, predicates in f, (s, p, o) triple patterns in f, or formal query language keywords used in f. If we have the following:

a. A set comparison function $g(f, f^*) \in [0, 1]$ that compares f and f^* as sets of elements $u_f \in \mathcal{U}_f$ and $u_{f^*} \in \mathcal{U}_{f^*}$, such that a correctly predicted formal query $f \overset{\text{sem}}{\equiv} f^* \implies g(f, f^*) = 1$.
b. Two predicted formal queries f_1 and f_2 where f_1 is a better prediction than f_2, i.e., where $1 > g(f_1, f^*) > g(f_2, f^*) \geq 0$.

Then we must also have that $m(f_1, f^*) > m(f_2, f^*)$.

Our second axiom (A2) corresponds to D2, concerning syntax correctness.

Axiom A2 - Executability: A predicted formal query must be syntactically well-formed and executable to be correct. If we have the following:

a. A ground truth formal query and its corresponding answer: $r_{\mathcal{K}}(f^*) = a^*$.
b. A predicted formal query f_1 that returns a non-error answer: $r_{\mathcal{K}}(f_1) = a_1 \neq \epsilon$.
c. A predicted formal query f_2 that results in an execution error: $r_{\mathcal{K}}(f_2) = \epsilon$.

Then we must have that $m(f_1, f^*) > m(f_2, f^*)$. This also holds if $a_1 = \emptyset \neq a^*$.

Note that this axiom assumes a strict definition of syntax correctness. If syntax correctness can be measured by degrees, then an additional axiom for D2 may be appropriate.

Our third axiom (A3) is corresponds to D3, concerning answer correctness.

Axiom A3 - Answer Completeness: Assuming an answer a as a set $\mathcal{T}_a$ of result tuples τ_a retrieved after executing a formal query f. If we have the following:

a. A set comparison function $g(a, a^*) \in [0, 1]$ that compares the predicted and ground truth answers as sets of elements $\tau_a \in \mathcal{T}_a$ and $\tau_{a^*} \in \mathcal{T}_{a^*}$, such that a correctly predicted answer $a = a^* \implies g(a, a^*) = 1$.
b. Two predicted answers a_1 and a_2 where a_1 is a better prediction than a_2, i.e., where $1 > g(a_1, a^*) > g(a_2, a^*) \geq 0$.

Then we must also have that $m(a_1, a^*) > m(a_2, a^*)$.

Table 1: Evaluating measures with axioms.

Measure	Axioms			General properties	
	A1	A2	A3	Instance level	Partial reward
Established measures					
Exact Match (Query)	◐[1]	◐	◐	●	○
Exact Match (Answer)	○	●	●	●	○
Acc/F_1/R/P (Answer)	○	●	●	●	●
Perplexity	○	○	○	●	●
BLEU	◐[1]	○	○	◐	●
ROUGE-L	◐[1]	○	○	●	●
Novel measures					
GEK-1	◐[1]	●	●	●	●
GEK-2	●	●	●	●	●
GEK-3	●	●	●	●	●

[1] Satisfies only A1.a.

4.2 Measure Analysis

Having established axioms derived from task-specific desiderata for evaluation measures, we can now analyze relevant established measures in terms of these axioms. We summarize our findings in Table 1, where ● indicates that a measure satisfies an axiom or general property, while ○ indicates it does not, and ◐ indicates a partial addressing of the axiom or property. Specifically, the general properties indicate whether a measure can be evaluated at an *instance level* and whether the measure can give a *partial reward* for a partially successful task.

We see that measures with n-gram-based matching like BLEU and ROUGE-L, as well as Exact Match (applied to the formal query) are able to address A1 partially, satisfying A1.a in that a perfect prediction will indeed give a maximal score of 1, assuming a single valid ground truth formal query for a given NL question. However, these measures do not capture specific semantic properties of the formal query, and so cannot fully satisfy A1. A major shortcoming shared by all the established measures is a failure to explicitly address whether a predicted formal query is executable. Assuming that the ground truth formal query is executable, which should be the case in principle, then Exact Match on the predicted formal query does satisfy A2 and A3. In practice, however, Query Exact Match by itself does not intrinsically guarantee A2 and A3, hence it can only partially satisfy those axioms.

Since D3 depends on D2, any evaluation of answer correctness assumes that the ground truth formal query has been correctly executed, which demonstrates syntax correctness D2. Hence, answer Exact Match and set-based measures applied only to answer correctness do imply the satisfaction of A2.

Exact Match cannot give partial reward, since only complete success is rewarded. We also note that while BLEU is not created to be evaluated on the instance level, it is possible to apply the measure in an instance-level manner. Finally, we note that Perplexity [14] does not satisfy any of the axioms even partially, and thus is not suitable for evaluating model performance for SP-KGQA. As can be seen from the analysis of measures against axioms, only our novel compound measures GEK-2 and GEK-3 fully satisfy all three axioms.

Table 2: Overview of transformations.

Original query				`SELECT DISTINCT ?uri where { dbr:Villa_Sturegården dbp:locationCountry ?uri }`
Trans.	**Desid.**			**Example degraded formal ground truth query (SPARQL)**
T_1	○	●	●	`SELECT DISTINCT ?uri where { dbr:Villa_Sturegården dbp:locationCountry ?uri`
T_2	●	○	●	`SELECT DISTINCT ?uri where { dbr:Yorkshire_1 dbp:champion ?uri }`
T_3	●	○	○	`SELECT DISTINCT ?uri where { dbr:Jonas_Kullhammar dbp:origin ?uri }`

Table 3: Synthetic experiments. Here † means moderately sensitive response ($\Delta > 0.10 \times x\%$ relative to T_0; ≤ 0.990 for $T_{i,10\%}$, and ≤ 0.980 $T_{i,20\%}$). ‡ means sensitive response ($\Delta > 0.50 \times x\%$ relative to T_0; $\leq$ 0.950 for $T_{i,10\%}$, and $\leq$0.900 for $T_{i,20\%}$).

Transf.	Established Measures				Novel Component Meas.				Novel Compound Meas.		
	Exact Match		BLEU	ROU.	Exec	$F_{1,Ans}$	$F_{1,Sem}$	$F_{1,Tri}$	GEK-1	GEK-2	GEK-3
	Query	Answer	Inst.								
T_0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$T_{1,10\%}$	$0.899^{\ddagger}$	$0.899^{\ddagger}$	$0.990^{\dagger}$	0.995	$0.899^{\ddagger}$	$0.899^{\ddagger}$	1.000	1.000	$0.899^{\ddagger}$	$0.899^{\ddagger}$	$0.899^{\ddagger}$
$T_{1,20\%}$	$0.799^{\ddagger}$	$0.799^{\ddagger}$	0.981	0.990	$0.799^{\ddagger}$	$0.799^{\ddagger}$	1.000	1.000	$0.799^{\ddagger}$	$0.799^{\ddagger}$	$0.799^{\ddagger}$
$T_{2,10\%}$	$0.900^{\ddagger}$	$0.900^{\ddagger}$	$0.929^{\ddagger}$	$0.962^{\dagger}$	$0.961^{\dagger}$	$0.900^{\ddagger}$	$0.900^{\ddagger}$	$0.900^{\ddagger}$	$0.900^{\ddagger}$	$0.900^{\ddagger}$	$0.900^{\ddagger}$
$T_{2,20\%}$	$0.799^{\ddagger}$	$0.800^{\ddagger}$	$0.859^{\ddagger}$	$0.925^{\dagger}$	$0.923^{\dagger}$	$0.800^{\ddagger}$	$0.799^{\ddagger}$	$0.799^{\ddagger}$	$0.800^{\ddagger}$	$0.799^{\ddagger}$	$0.799^{\ddagger}$
$T_{3,10\%}$	$0.942^{\ddagger}$	1.000	$0.965^{\dagger}$	0.983	1.000	1.000	$0.957^{\dagger}$	$0.948^{\ddagger}$	$0.965^{\ddagger}$	$0.957^{\ddagger}$	$0.948^{\ddagger}$
$T_{3,20\%}$	$0.885^{\ddagger}$	1.000	$0.931^{\ddagger}$	$0.966^{\ddagger}$	1.000	1.000	$0.915^{\ddagger}$	$0.898^{\ddagger}$	$0.931^{\dagger}$	$0.915^{\dagger}$	$0.898^{\ddagger}$

5 EXPERIMENTS

We investigate the proposed SP-KGQA framework, and more specifically GEK-1..3 and established measures, with two experiments. First, we investigate the sensitivity of measures in Sect. 5.2 by evaluating *synthetic runs* with controlled degradation of the ground truth. Second, we evaluate neural models trained on a large complex SP-KGQA dataset in Sect. 5.3.

5.1 Experimental Setup

We use the DBNQA [30] dataset, which consists of 894,499 instances, generated from 5,217 templates extracted from the LC-QuAD [59] and QALD-7-Train [61] datasets. Specifically, to avoid the reported information leakage issues [35], a sanitized partitioning of DBNQA is used, here referred to as DBNQA* (called Sanitized-1 in [35]), where the training and test splits are partitioned based on the underlying templates. In terms of the Chakraborty et al. [18] categories for neural KGQA, our work focuses on translation-based SP-KGQA, and our setting is fully supervised machine learning.

To retrieve answers, all formal queries, including the ground truth, those predicted by models, and those generated for the synthetic runs, are executed against a Virtuoso endpoint holding DBpedia 2016-10 [12] as the KG. The retrieved results are considered the respective ground truth and predicted answers.

During training, prediction, and query evaluation, the formal queries were tokenized and encoded in the manner of Soru et al. [55, 56]. We refer to Yin et al. [65] for an encoding example. For calculating GEK-1..3, we set the value $\gamma = 10^{-4}$.

5.2 Experiment 1: Synthetic Experiments

To investigate the sensitivity of SP-KGQA measures, we simulate predictions errors without training models by constructing *synthetic runs* by degrading ground truth test data in a controlled manner. The more degradation applied, the worse the synthetic predictions; consequently, the more evaluation scores should decrease. This demonstrates that our novel compound measures are sensitive and balanced to all of the tested error types, as compared to established measures.

5.2.1 Transformations. We devise a set of *transformations*, that simulate particular types of prediction errors. Each of these transformations preserve the prediction quality with respect to one desideratum while degrading with regards to others. Table 2 provides an overview of the transformations, where desiderata marked with ○ are preserved, while those with ● are disrupted. Each transformation is randomly applied to 10% or 20% of the test split.

- T_1: Remove the closing curly bracket (}) from the `WHERE` clause. This mostly preserves D1 (both *n*-gram overlaps and semantic properties, like URIs and SPO triple patterns), but disrupts D2, and hence D3. This creates the case of correct semantics but wrong result.
- T_2: Replace each entity (predicate) URI with a random entity (predicate) URI. This mostly preserves D2, but deteriorates D1 and hence D3.
- T_3: Replace query with another that yields the exact same answer, if another such query exists in the dataset. This preserves D2 and D3, but disrupts D1.

5.2.2 Results. Table 3 presents the results of the synthetic experiments, where results meeting the sensitivity thresholds are indicated by daggers.[2] For each transformation T_i and for each measure, we see that established measures, novel component measures, and novel compound measures are reduced either proportionally by the degradation, or else negligibly. For example, T_1 affects Exact

[2]We omit corpus BLEU because results correlate closely with instance-level BLEU.

Table 4: Evaluation of SP-KGQA methods in terms of established and novel measures. Best scores in each block are boldfaced.

Method	Training data	Established Measures					Novel Compound Meas.		
		Exact Match		BLEU		ROUGE	GEK-1	GEK-2	GEK-3
		Query	Answer	Corpus	Instance				
NSpM	12.5%	**0.000**	**0.059**	**0.437**	**0.374**	**0.679**	**0.009**	**0.006**	0.002
	25%	**0.000**	0.034	0.434	0.371	0.676	**0.009**	**0.006**	**0.003**
	50%	**0.000**	0.019	0.432	0.371	0.678	0.006	0.002	0.001
	100%	**0.000**	0.023	0.432	0.371	**0.679**	0.006	0.004	0.001
NSpM+Att1	12.5%	0.012	0.036	0.486	0.417	0.713	0.022	0.022	0.017
	25%	0.024	0.050	0.484	0.423	0.721	0.037	0.035	0.029
	50%	0.045	0.074	0.511	0.451	0.740	0.060	0.060	0.052
	100%	**0.081**	**0.117**	**0.548**	**0.498**	**0.778**	**0.105**	**0.105**	**0.093**
NSpM+Att2	12.5%	0.008	0.036	0.478	0.408	0.705	0.017	0.016	0.012
	25%	0.029	0.053	0.498	0.436	0.731	0.041	0.041	0.035
	50%	0.049	0.076	0.522	0.460	0.747	0.066	0.066	0.058
	100%	**0.078**	**0.119**	**0.558**	**0.503**	**0.785**	**0.107**	**0.107**	**0.093**
ConvS2S	12.5%	0.042	0.062	0.485	0.444	0.759	0.056	0.056	0.050
	25%	0.066	0.106	0.536	0.490	0.795	0.094	0.100	0.089
	50%	0.084	0.119	0.577	0.519	0.817	0.110	0.112	**0.100**
	100%	**0.085**	**0.126**	**0.582**	**0.525**	**0.821**	**0.115**	**0.114**	**0.100**
Transformer	12.5%	0.051	0.089	0.489	0.436	0.764	0.082	0.088	0.073
	25%	0.077	0.154	0.528	0.480	0.791	0.137	0.147	0.123
	50%	0.102	0.202	0.560	0.510	0.809	0.179	0.193	0.163
	100%	**0.113**	**0.229**	**0.570**	**0.522**	**0.810**	**0.199**	**0.217**	**0.180**

Match proportionally at a one-to-one rate, while BLEU and ROUGE-L reduce at a lower rate. This illustrates the independence of D1 from D2 and D3. In contrast, $F_{1,Sem}$ and $F_{1,Tri}$ are not affected by T_1. Crucially, we see that GEK-1..3 and Query Exact Match are sensitive to all transformations. Out of these measures, GEK-3 is the only measure considered here that satisfies all the axioms and also shows sensitivity to all the synthetic prediction errors tested. Therefore, if a single measure is to be used, we recommend that to be GEK-3.

5.3 Experiment 2: Neural Methods

So far, we have shown both theoretically (with our axiomatic analysis in Sect. 4.1) and empirically (with our controlled degradation experiments in Sect. 5.2) that our novel GEK-1..3 measures indeed capture the desired properties of the SP-KGQA task. Having a principled and validated measurement instrumentation at our disposal, we are interested in applying the novel measures to evaluate state-of-the-art NMT SP-KGQA models, and assess whether the findings agree with the reported results in previous works using established measures.

5.3.1 Methods.

- **NSpM [55, 56]:** architectures Baseline, Attention 1, and Attention 2 were based on the original Tensorflow NMT implementation. Hyperparameters were 50 000 training steps, 2 layers, dropout 0.2. Based on [65], Attention 1 used the normed Bahdanau [8] attention mechanism, while Attention 2 used the scaled Luong [37] attention mechanism.
- **ConvS2S [29]:** architecture based on PyTorch implementation,[3] using default hyperparameters, except optimizing with stochastic gradient descent with learning rate of 0.075 and momentum of 0.99.
- **Transformer [63]:** architecture based on the same PyTorch project as ConvS2S, using only default hyperparameters.

Some NMT models perform better with sub-word tokenization, such as Byte Pair Encoding (BPE) [52], so for the Transformer and ConvS2S models we additionally used Sentencepiece[4] BPE restricted to maximum 32k terms. Increased volumes of training data are expected to produce better performance if the model is learning. We train models from each of the five architectures on four different training splits from DBNQA*: 12.5%, 25%, 50%, and 100%. We can then compare the measures on models expected to show performance improvement.

5.3.2 Results. Table 4 presents the results of SP-KGQA neural models trained on different training data volumes. The NSpM baseline models show negligible change in all measures; this indicates that despite the additional training instances no improvement has occurred. All other methods show improvements in all measures with increased training data. Note that different models improve more on some measures than others. For example, Transformer is not the best model with respect to BLEU or ROUGE-L, but is clearly the best with respect to Query and Answer Exact Match.

[3] https://github.com/bentrevett/pytorch-seq2seq/
[4] https://github.com/google/sentencepiece

As Table 4 shows, comparing Exact Match for queries and answers, there can be a large gap between semantic parsing quality and answer correctness. Syntax correctness also shows an SP-KGQA system's mastery of the formal query syntax (e.g., SPARQL). Therefore, our novel compound measures give partial reward to models which generate SPARQL queries which are executable without any errors.

Yin et al. [65] concluded that "the ConvS2S model consistently, significantly outperformed all other models [including Transformer] at a margin," but our results constitute evidence to the contrary. The change between the relative ordering of ConvS2S and Transformer is a crucial one, especially considering the absolute performance difference. While the two are close in terms of BLEU and ROUGE-L (with a slight advantage to ConvS2S), the Transformer model produces substantially more executable queries (see Answer Exact Match scores). Hence, measures of individual components in the SP-KGQA task may give a distorted impression of the system performance on the task as a whole. This shows the advantage of a compound measure being simultaneously sensitive to all desiderata.

6 DISCUSSION AND CONCLUSION

After first discussing limitations and future directions, we conclude with our key steps and insights.

6.1 Limitations and Future Directions

Being a first study in this direction, the work presented is not without limitations.

6.1.1 An Initial Axiomatic Effort. We identified the desiderata based on the essential qualities of the KGQA task [18], following a standard methodology of axiomatic development of evaluation measures in Sect. 2.2. We acknowledge that our desiderata and axioms represent one possible perspective and it is by no means claimed to be exhaustive. For example, the answer completeness axiom makes strong assumptions about how a retrieved answer is to be treated, i.e., as a set of result tuples. This restricted view of answer correctness does not consider the relationships between the elements of an answer. In the future, additional or alternative axiom-level formulations of answer correctness may be developed, which would regard it as degrees of plausibility, e.g., with respect to the type of the answer retrieved [41]. Further desiderata may be included, e.g., to express a preference for simpler and shorter predicted formal queries. Syntax correctness could also be extended to address degrees of partial correctness. Furthermore, we plan to further expand the set of axioms and component measures to consider structural elements of semantic representation, like reserved words, brackets, and query graph structure.

6.1.2 Choice of Component Measures. In the present work, BLEU and ROUGE are interpreted as reflecting semantic representation because they would partly capture this aspect, at least matching the entity and predicate URI unigrams in a formal query. As for the novel component measures of semantic representation, entities and predicates are treated as more important semantic signifiers than structural elements of the formal queries because they are the explicit KG properties which must be matched. However, the structural elements play an important semantic role, and future work could incorporate these.

Validating whether a formal query provides a good semantic representation of a given NL question would require expert human evaluation efforts. This is similar to the challenge that motivated measures such as BLEU and ROUGE initially. We introduce our overlap-based measures as a first effort at automating the evaluation in terms of salient semantic elements from the formal query. Future work can investigate which semantic representation measures might correlate best with human evaluations of semantic representation. Importantly, the key message of our paper is not about the specific component measures suggested, but about the framework of taking multiple desiderata of a single task into account simultaneously in a formally grounded manner.

6.1.3 Weighting of Component Measures. The component measures corresponding to the three task desiderata are equally weighted in our proposed compound measure framework because all the desiderata are necessary for the SP-KGQA task. There is no *a priori* difference in importance between the three desiderata. However, as required for a particular system or application scenario, our framework enables different weightings of component measures.

6.1.4 Generalizing to Multiple Knowledge Graphs. We have experimentally addressed KGQA on a single, well-maintained ontology, DBpedia. Since this is an initial iteration on an axiomatic approach to developing grounded evaluation measures for SP-KGQA, we have restricted the scope of the work to general aspects of KGs and formal queries. This has been a helpful constraint to be able to clearly express our framework. However, in future work it may be interesting to develop concepts in our framework with respect to semantic knowledge representation, e.g., as expressed using the OWL Web Ontology Language. Distinct URIs may represent the same underlying property or entity, in which case they are connected by the OWL predicate `owl:sameAs`. If two predicates represent mutually inverse properties, the are connected by `owl:inverseOf`.

Extending the first NL question example in Sect. 3.1, with a pair of predicates that are mutual inverses such as `parentOf` and `childOf`, we can replace the triple pattern `?person childOf ?parent` with `?parent parentOf ?person`. We then get yet another formal query formulation

```
SELECT ?person WHERE { ?parent parentOf ?person
 . ?parent childOf Albert_Einstein }
```

that we can recognize as semantically equivalent to the two formal queries in the Sect. 3.1 example, such that all three formal queries correctly represent the meaning of the NL question "Who are the grandchildren of Albert Einstein?" This illustrates how semantic technologies, like OWL, may be used to expand the concept of semantic equivalence. Developments in this direction would also enable our framework to evaluate QA approaches over multiple KGs. The complexity of evaluating SPARQL semantic representations in a multi-KG ontology may be addressed by resolving the synonymy of different URIs for the same entity or predicate.

6.2 Conclusion

Currently, there is no agreed-upon way of evaluating SP-KGQA systems. Previous work uses multiple measures that evaluate individual aspects, such as either the quality of the semantic parse or the quality of the answer retrieved. It is clear that researchers also want to consider multiple aspects, but there is no other way of doing that other than reporting on a set of different measures. There is no systematic and principled way to combine these aspects in a single unified evaluation measure. Because of that, one particular measure (implicitly or explicitly) becomes "privileged" and gets optimized, at the expense of others. This carries the risk of overfitting and may lead to an imbalanced view of true system performance. It is therefore clear that there is a need to unify SP-KGQA measures in a formally grounded manner.

We have looked through an axiomatic lens at the measures used to evaluate systems' performance on the SP-KGQA task. We have introduced a probabilistic framework for a family of compound measures capable of addressing all the identified task desiderata. With this framework we have instantiated novel compound measures, designed specifically for the SP-KGQA task. After postulating axioms for SP-KGQA evaluation measures, our axiomatic analysis found insufficiencies in established measures that our novel compound measures resolve. We have also validated the novel measures by evaluating synthetic predictions, before evaluating real predictions by state-of-the-art NMT-based SP-KGQA models. From the experiments, we see that established measures are generally sensitive to some but not all desiderata, unlike our novel compound measures. The discrepancy between established and novel measures we have observed in state-of-the-art NMT-based SP-KGQA models indicates a need for re-examination of the results of previous works.

REFERENCES

[1] Ameer Albahem, Damiano Spina, Falk Scholer, Alistair Moffat, and Lawrence Cavedon. 2018. Desirable Properties for Diversity and Truncated Effectiveness Metrics. In *Proceedings of the 23rd Australasian Document Computing Symposium (ADCS '18)*.

[2] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. 2017. Axiomatic Thinking for Information Retrieval: And Related Tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 1419–1420.

[3] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints. *Information Retrieval Journal* 12, 4 (2009), 461–486.

[4] Enrique Amigó, Julio Gonzalo, Felisa Verdejo, and Damiano Spina. 2019. A Comparison of Filtering Evaluation Metrics Based on Formal Constraints. *Information Retrieval Journal* 22, 6 (2019), 581–619.

[5] Enrique Amigó and Stefano Mizzaro. 2020. On the Nature of Information Access Evaluation Metrics: A Unifying Framework. *Information Retrieval Journal* 23, 3 (2020), 581–619.

[6] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 625–634.

[7] Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic Parsing as Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL '13)*. 47–52.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs.CL]

[9] Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based Question Answering with Knowledge Graph. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (COLING '16)*. 2503–2514.

[10] Hannah Bast and Elmar Haussmann. 2015. More Accurate Question Answering on Freebase. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. 1431–1440.

[11] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP '13)*. 1533–1544.

[12] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics* 7, 3 (2009), 154–165.

[13] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale Simple Question Answering with Memory Networks. arXiv:1506.02075 [cs.CL]

[14] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. 1992. An Estimate of an Upper Bound for the Entropy of English. *Computational Linguistics* 18, 1 (1992), 31–40.

[15] Luca Busin and Stefano Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval (ICTIR '13)*. 22–29.

[16] Jianyu Cai, Zhanqiu Zhang, Feng Wu, and Jie Wang. 2021. Deep Cognitive Reasoning Network for Multi-Hop Question Answering over Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 219–229.

[17] Qingqing Cai and Alexander Yates. 2013. Large-Scale Semantic Parsing via Schema Matching and Lexicon Extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '13)*. 423–433.

[18] Nilesh Chakraborty, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann, and Asja Fischer. 2021. Introduction to Neural Network-based Question Answering over Knowledge Graphs. *WIREs Data Mining and Knowledge Discovery* 11, 3 (2021), e1389.

[19] Dennis Diefenbach, José Giménez-García, Andreas Both, Kamal Singh, and Pierre Maret. 2020. QAnswer KG: Designing a Portable Question Answering System over RDF Data. In *Proceedings of the 2020 European Semantic Web Conference: The Semantic Web (ESWC '20)*. 429–445.

[20] Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core Techniques of Question Answering Systems over Knowledge Bases: A Survey. *Knowledge and Information Systems* 55, 3 (2018), 529–569.

[21] Dennis Diefenbach, Thomas Tanon, Kamal Singh, and Pierre Maret. 2017. Question Answering Benchmarks for Wikidata. In *Proceedings of the 16th International Semantic Web conference (ISWC '17)*.

[22] Jiwei Ding, Wei Hu, Qixin Xu, and Yuzhong 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 '19)*. 2614–2622.

[23] Li Dong and Mirella Lapata. 2016. Language to Logical Form with Neural Attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL '16)*. 33–43.

[24] Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *Proceedings of the 18th International Semantic Web Conference (ISWC '19)*. 69–78.

[25] Marco Ferrante, Nicola Ferro, and Eleonora Losiouk. 2020. How Do Interval Scales Help Us with Better Understanding IR Evaluation Measures? *Information Retrieval Journal* 23, 3 (2020), 289–317.

[26] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2015. Towards a Formal Framework for Utility-Oriented Measurements of Retrieval Effectiveness. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval (ICTIR '15)*. 21–30.

[27] Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. 2017. Are IR Evaluation Measures on an Interval Scale?. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '17)*. 67–74.

[28] Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. 2019. A General Theory of IR Evaluation Measures. *IEEE Transactions on Knowledge and Data Engineering* 31, 3 (2019), 409–422.

[29] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. 1243–1252.

[30] Ann-Kathrin Hartmann, Edgard Marx, and Tommaso Soru. 2018. Generating a Large Dataset for Neural Question Answering over the DBpedia Knowledge Base. In *Workshop on Linked Data Management, co-located with the W3C WEBBR (W3C WEBBR '18)*.

[31] Xin Hu, Jiangli Duan, and Depeng Dang. 2021. Natural Language Question Answering over Knowledge Graph: The Marriage of SPARQL Query and Keyword Search. *Knowledge and Information Systems* 63, 4 (2021), 819–844.

[32] Edwin T. Jaynes. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

[33] Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21 (IJCAI '21)*. 4483–4491.

[34] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. 74–81.

[35] Trond Linjordet and Krisztian Balog. 2020. Sanitizing Synthetic Training Data Generation for Question Answering over Knowledge Graphs. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR '20)*. 121–128.

[36] Kangqi Luo, Fengli Lin, Xusheng Luo, and Kenny Zhu. 2018. Knowledge Base Question Answering via Encoding of Complex Query Graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. 2185–2194.

[37] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. arXiv:1508.04025 [cs.CL]

[38] Gaurav Maheshwari, Priyansh Trivedi, Denis Lukovnikov, Nilesh Chakraborty, Asja Fischer, and Jens Lehmann. 2019. Learning to Rank Query Graphs for Complex Question Answering over Knowledge Graphs. In *Proceedings of the 18th International Semantic Web Conference*. 487–504.

[39] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[40] Raymond J Mooney. 2007. Learning for Semantic Parsing. In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing' 07)*. 311–324.

[41] Christos Nikas, Pavlos Fafalios, and Yannis Tzitzikas. 2021. Open Domain Question Answering over Knowledge Graphs Using Keyword Search, Answer Type Prediction, SPARQL and Pre-trained Neural Models. In *Proceedings of the 20th International Semantic Web Conference (ISWC '21)*. 235–251.

[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*. 311–318.

[43] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Proceedings of ACM Conference on Recommender Systems (RecSys '21)*.

[44] Michael Petrochuk and Luke Zettlemoyer. 2018. SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*. 554–558.

[45] Joan Plepi, Endri Kacupaj, Kuldeep Singh, Harsh Thakkar, and Jens Lehmann. 2021. Context Transformer with Stacked Pointer Networks for Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 2021 European Semantic Web Conference: The Semantic Web (ESWC '21)*. 356–371.

[46] John Prager. 2007. Open-Domain Question-Answering. *Foundations and Trends in Information Retrieval* 1, 2 (2007), 91–231.

[47] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

[48] Rishiraj Saha Roy and Avishek Anand. 2021. Question Answering for the Curated Web: Tasks and Methods in QA over Knowledge Bases and Text Collections. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 13, 4 (2021), 1–194.

[49] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 595–604.

[50] Fabrizio Sebastiani. 2015. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. 11–20.

[51] Fabrizio Sebastiani. 2020. Evaluation Measures for Quantification: an Axiomatic Approach. *Information Retrieval Journal* 23, 3 (2020), 255–288.

[52] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. arXiv:1508.07909 [cs.CL]

[53] Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base. 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 '19)*. 2442–2451.

[54] Kuldeep Singh, Ioanna Lytra, Arun Sethupat Radhakrishna, Saeedeh Shekarpour, Maria-Esther Vidal, and Jens Lehmann. 2020. No One Is Perfect: Analysing the Performance of Question Answering Components over the DBpedia Knowledge Graph. *Journal of Web Semantics* 65 (2020).

[55] Tommaso Soru, Edgard Marx, Diego Moussallem, Gustavo Publio, André Valdestilhas, Diego Esteves, and Ciro Baron Neto. 2017. SPARQL as a Foreign Language. arXiv:1708.07624 [cs.CL]

[56] Tommaso Soru, Edgard Marx, André Valdestilhas, Diego Esteves, Diego Moussallem, and Gustavo Publio. 2018. Neural Machine Translation for Query Construction and Composition. In *ICML Workshop on Neural Abstract Machines & Program Induction (NAMPI v2) (ICML '18)*.

[57] Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gür, Zenghui Yan, and Xifeng Yan. 2016. On Generating Characteristic-Rich Question Sets for QA evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '26)*. 562–572.

[58] Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (NAACL-HLT '18)*. 641–651.

[59] Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In *Proceedings of the 16th International Semantic Web Conference (ISWC '17)*. 210–218.

[60] Christina Unger, André Freitas, and Philipp Cimiano. 2014. *An Introduction to Question Answering over Linked Data*. 100–140.

[61] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th Open Challenge on Question Answering over Linked Data (QALD-7). In *Semantic Web Evaluation Challenge (SWEC '17)*. 59–69.

[62] Ricardo Usbeck, Michael Röder, Michael Hoffmann, Felix Conrads, Jonathan Huthmann, Axel-Cyrille Ngonga-Ngomo, Christian Demmler, and Christina Unger. 2019. Benchmarking Question Answering Systems. *Semantic Web* 10, 2 (2019), 293–304.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS'17)*. 6000–6010.

[64] Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP '15)*. 1321–1331.

[65] Xiaoyu Yin, Dagmar Gromann, and Sebastian Rudolph. 2021. Neural Machine Translating from Natural Language to SPARQL. *Future Generation Computer Systems* 117 (2021), 510–519.