



Completeness, Recall, and Negation in Open-world Knowledge Bases: A Survey

SIMON RAZNIEWSKI, Bosch Center for AI, Germany

HIBA ARNAOUT, TU Darmstadt, Germany

SHRESTHA GHOSH, Max Planck Institute for Informatics, Germany

FABIAN SUCHANEK, Telecom Paris, Institut Polytechnique de Paris, France

General-purpose knowledge bases (KBs) are a cornerstone of knowledge-centric AI. Many of them are constructed pragmatically from web sources and are thus far from complete. This poses challenges for the consumption as well as the curation of their content. While several surveys target the problem of completing incomplete KBs, the first problem is arguably to know whether and where the KB is incomplete in the first place, and to which degree.

In this survey, we discuss how knowledge about completeness, recall, and negation in KBs can be expressed, extracted, and inferred. We cover (i) the logical foundations of knowledge representation and querying under partial closed-world semantics; (ii) the estimation of this information via statistical patterns; (iii) the extraction of information about recall from KBs and text; (iv) the identification of interesting negative statements; and (v) relaxed notions of relative recall.

This survey is targeted at two types of audiences: (1) practitioners who are interested in tracking KB quality, focusing extraction efforts, and building quality-aware downstream applications; and (2) data management, knowledge base, and semantic web researchers who wish to understand the state-of-the-art of knowledge bases beyond the open-world assumption. Consequently, our survey presents both fundamental methodologies and the results that they have produced, and gives practice-oriented recommendations on how to choose between different approaches for a problem at hand.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Knowledge representation and reasoning**; **Artificial intelligence**;

Additional Key Words and Phrases: Knowledge bases, data completeness

ACM Reference Format:

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, and Fabian Suchanek. 2024. Completeness, Recall, and Negation in Open-world Knowledge Bases: A Survey. *ACM Comput. Surv.* 56, 6, Article 150 (February 2024), 42 pages. <https://doi.org/10.1145/3639563>

Authors' addresses: S. Razniewski, Bosch Center for AI, Robert-Bosch-Campus 1, 71272 Renningen, Germany; e-mail: simon.rzniewski@de.bosch.com; H. Arnaout, TU Darmstadt, Hochschulstraße 10, 64289 Darmstadt, Germany; e-mail: harnaout@mpi-inf.mpg.de; S. Ghosh, Max Planck Institute for Informatics, Germany; e-mail: ghoshs@mpi-inf.mpg.de; F. Suchanek, Telecom Paris, Institut Polytechnique de Paris, 19 place Marguerite Perey, 91120 Palaiseau, France; e-mail: suchanek@telecom-paris.fr.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 0360-0300/2024/02-ART150

<https://doi.org/10.1145/3639563>

1 INTRODUCTION

Motivation. Web-scale **knowledge bases (KBs)** such as Wikidata [125], DBpedia [13], NELL [21], or YAGO [118] are a cornerstone of the Semantic Web. Pragmatically constructed from web resources, they focus on representing *positive* knowledge, i.e., statements that are true. However, they typically contain only a small subset of all true statements, without qualifying what that subset is. For example, a KB may contain winners of the *Nobel Prize in Physics*, but it does not necessarily contain *all* winners. It will not even specify whether it contains all winners or not. For example, a KB may lack the information that a specific renowned physicist won the *Nobel Prize*, but this does not mean that this person did not win the award—the data may just be lacking. Vice versa, if it is known that a specific physicist did definitively not win the *Nobel Prize*, then there is no way to express this in most current KBs. A KB may also not contain a specific scientist at all—without any indication that she or he is missing.¹

Such uncertainty about the extent of knowledge poses major challenges for the curation and application of KBs:

- (1) **Human KB curators** need to know where the KB is incomplete so they can prioritize their completion efforts. For example, when working on a KB like Wikidata, with over 100M items, systematic knowledge on gaps and quality is essential to know where to focus limited human resources.
- (2) **Automated KB construction pipelines** need this knowledge, too, to know how to adjust acceptance thresholds. For example, if we already have 218 Physics Nobel prize winners in the KB, and if we know that this is the expected count, then further candidates should all be rejected. In contrast, if no publications of a scientist are recorded in the KB, then we should accept automated extractions. This is significant in particular for KB projects such as NELL, which aim to auto-complete themselves.
- (3) **QA Applications** are built on top of KBs. They need to know where the data is incomplete, to alert end-users of quality issues. For example, a query for “*the astronomer who discovered most planets*” may return the wrong answer if *Geoffrey Marcy* happens to be absent from the KB. Similarly, a KB that is used for question-answering should have awareness of when a question surpasses its knowledge [93]. This holds in particular for Boolean questions such as “*Does Harvard have a physics department?*” (where a “no” could come simply from missing information) and analytical aggregate questions such as “*How many US universities have physics departments?*” (where receiving some answer gives little clue about its correctness).
- (4) **Structured entity summaries** need awareness not only of positive properties, but also of salient negatives. For example, one of the most notable properties of *Switzerland* is that it is not a *EU* member (despite heavy economic ties and being surrounded by EU countries). For *Stephen Hawking*, a salient summary of his accolades should include that he did *not* win the #1 award in his field, the *Nobel Prize in Physics*.
- (5) **Machine learning** on knowledge bases and text, in particular, for the tasks of KB link prediction and textual information extraction, needs negative training examples to build accurate models. Obtaining quality negative examples is a major hurdle when working on these tasks, and much research has focused on heuristics, for example, based on obfuscation of positive statements. Explicit negations, or negations derived from completeness metadata, could significantly impact these tasks.

This survey presents the methods that the recent literature has developed to address these problems. For example, several formalisms have been developed to specify the extent of the

¹We refrain from giving an example of an incomplete entity in a real-world KB, because whenever we did that in previous publications, helpful readers specifically completed these entities in the KB, thereby rendering our example outdated.

knowledge in a KB, including the possibility to make negative statements or to specify metadata on completeness. Knowledge engineers can record that they entered *all* winners of the *Nobel Prize in Physics* into Wikidata or mark prominent Physicists who did *not* win it. If such a manual marking is not possible, then there are also methods to infer the completeness of a KB automatically. For example, it is possible to spot phrases such as “*the Nobel Prize in Physics was awarded to 218 individuals*” and to compare this count with the number of entities in the KB. It is also possible to draw conclusions about the completeness of a KB from the growth or near-stagnation of a set of entities in the KB, from the overlap between sets of *Nobel Prize* winners from different random samples or from different KBs. One can also observe that most winners hold academic degrees and so flag entities without an alma mater as likely incomplete. The knowledge engineer thus has a growing repository of methods at hand to tackle the problem of incompleteness, and this is what the present survey covers.

Focus of this survey. Several comprehensive surveys discuss KBs at large, in particular their construction, storage, querying, and applications [50, 55, 127]. Focused works on KB quality discuss how recall can be increased and how quality can be measured by assessing accuracy and provenance [17, 35, 85, 130]. Recent years have also seen new formalisms for describing recall and negative knowledge [10, 26, 71], as well as the rise of statistical and text-based methods for estimating recall [15, 36, 52, 58, 61, 67, 97, 114] and deriving negative statements [8, 12, 58, 108], with some of them collected in a systematic literature review [53]. The goal of this survey is to systematize, present, and analyze this topic. The survey is intended both for theoreticians and practitioners. It intends to inform readers about the latest advances in completeness assessment and negation and equip them with a repertoire of methodologies to better represent and assess the recall of specific datasets. The survey builds on content that has been presented at tutorials at VLDB’21, ISWC’21, KR’21, and WWW’22. The tutorial slides are available at <https://www.mpi-inf.mpg.de/knowledge-base-recall/tutorials>

Outline. This survey is structured as follows: In Section 2, we start with the **foundations** of the logical framework in which KBs operate, the **open world assumption (OWA)**, incompleteness, the implications for query answering [102], as well as the formal semantics of completeness and cardinality assertions [26]. In Section 3, we discuss **how recall can be estimated automatically**. We present supervised machine learning methods [36], unsupervised techniques such as species sampling [67, 123], density-based estimators [61], statistical invariants about number distributions [114], and linguistic theories about completeness in human conversations [97]. In Section 4, we focus on **determining the cardinality of a predicate** (i.e., its number of objects for a given subject) from KBs and from text. We show how cardinality information from KBs can be identified and used to assess recall [42], how this information can be extracted from natural language documents [71], and how this differs from mining functional cardinalities that remain invariant given any subject (for instance, all humans have one birthplace) [43]. In Section 5, we focus on **identifying salient negations**. We show why explicit negations about KB entities are needed in open-world settings. We present methods to identify negations using inferences from the KB itself [8, 11, 12, 107, 108] and methods to extract negations from various textual sources [8, 58], in particular, query logs. We also outline open issues such as the precision/salience tradeoff and ontology modeling and maintenance. In Section 6, we discuss **relative recall**, i.e., more relaxed notions of recall that stop short of aiming to capture all knowledge from the real world. We show how recall can be measured by extrinsic use cases such as question answering and entity summarization [52, 96], by comparison with open information extraction or external reference resources [39, 73] and by comparison with other comparable entities inside the KB [15, 70]. In Section 7, we conclude with **recent topics, recommendations** towards making KBs recall-aware, and **open research**

challenges. In particular, **large language models (LLMs)** have recently shaken up knowledge-centric computing. Although completeness and recall research has yet to capitalize on these advances, we highlight several ways by which LLMs are likely to impact this area.

2 FOUNDATIONS

2.1 Knowledge Bases

We first introduce foundational concepts, in particular, knowledge bases, their semantics, and the notions of completeness, recall, and cardinality metadata. Knowledge bases are built on three pairwise disjoint infinite sets E (entities), P (predicates), and L (literals). Entities refer to uniquely identifiable objects such as *Marie Curie*, the *Nobel Prize in Physics*, or the city of *Paris*. Literals are strings or numerical values, such as *dates*, *weights*, and *names*. Predicates (also known as relations, properties, or attributes) link an entity to another entity or to a literal. Examples are *birthPlace* or *hasAtomicNumber*. A tuple $\langle s, p, o \rangle \in E \times P \times (E \cup L)$ is called a **positive statement** (also known as triple, assertion, or fact), where s is the subject, p the predicate, and o the object [119]. A statement says that the subject stands in the given predicate with the object, as in $\langle \textit{Marie Curie}, \textit{wonAward}, \textit{Nobel Prize in Physics} \rangle$. For our purposes, we will consider also **negative statements**, which say that the subject does not stand in the given relationship with the object, as in $\neg \langle \textit{Stephen Hawking}, \textit{wonAward}, \textit{Nobel Prize in Physics} \rangle$. Entities can be organized into classes such as *Physicists*, *Cities*, or *Awards*. An entity can be an *instance* of a class, and this can be expressed by a triple with the predicate *type*, as in $\langle \textit{Marie Curie}, \textit{type}, \textit{Physicist} \rangle$. Classes can be arranged in a subclass hierarchy. For example, physicists are scientists, and scientists are people. This, too, can be expressed by triples, using the special predicate *subClassOf*: $\langle \textit{Physicist}, \textit{subClassOf}, \textit{Scientist} \rangle$. Some KBs allow specifying *constraints*, such as domain and range constraints on predicates, functionality constraints, or disjointness constraints between classes, usually in a formalism called Description Logic [14]. In our case, we are concerned mainly with the statements and with class membership and not so much with constraints (i.e., in the terminology of Description Logic, we focus on the assertional part of knowledge of the A-box and less on the terminological knowledge of the T-box).

Definition 1 (Knowledge Base). A knowledge base (or knowledge graph) K is a finite set of positive statements [60].

Example: As a running example, consider a KB with biographical data about scientists. An excerpt of this KB is shown in Table 1, in a Wikidata-style layout, where the subject of all statements is shown at the top (*Marie Curie*), and predicates and objects are shown in tabular form below. This KB contains a diverse set of statements, linking entities with other entities, as well as with literals, and containing multiple objects for *some* predicates but not for *all*.

The term **Knowledge Graph (KG)** is also often used as an alternative to “Knowledge Base”—ever since Google popularized the term in 2012 [110]. Although KGs are sometimes defined as special cases of KBs, by and large, the two terms have been used synonymously. For the purpose of this survey, we follow suit and use them interchangeably. KBs (or KGs) have received a considerable uptake in recent years [50, 127]. From humble beginnings in dedicated communities such as YAGO [118], DBpedia [13], Freebase [20], or Wikidata [125], KBs are now standard technology at most major tech companies [80] and widely in use beyond. Nonetheless, understanding their quality remains an enduring challenge. With this survey, we contribute a detailed overview concerning the aspects of completeness, recall, and negation.

In this survey, we focus on knowledge that can be found in text and that can reasonably be expressed in structured format in knowledge bases. Intangible knowledge, e.g., around concepts such as motion, scent, rhythm, or touch, is out of scope for this article.

Table 1. Running Example KB

Marie Curie	
birthPlace	Warsaw
birthYear	1867
citizenOf	Poland
advised	Marguerite Perey Óscar Moreno ... (6 more names) ...
discovered	Polonium
wonAward	Nobel Prize in Physics Franklin Medal

2.2 Incompleteness

Open-world knowledge bases are inherently incomplete, i.e., they store only a subset of all true statements from the domain of interest [100]. Our example KB in Table 1, for instance, misses two other citizenships that Marie Curie held (*Russia, France*), and it does not even contain her receiving the *Nobel Prize in Chemistry*.² Moreover, it lacks the statement that, contrary to a common misconception, she did not find the first discovered radioactive element *Uranium*.

Such incompleteness happens for several reasons:

- (1) **Source incompleteness:** KBs are usually built from another source, either by automated methods [127] or by human curators, who read websites, textbooks, news articles, and so on, or other authoritative websites. These sources may themselves be incomplete.
- (2) **Access limitations:** In some cases, only a subset of the relevant documents is readily accessible to automated methods or to human curators. For example, major troves of knowledge are locked away in the Deep Web (web pages without inlinks) [19], and there are printed documents that are not available in digital form.
- (3) **Extraction imperfection:** Even statements that are available in digital form and accessible can be missed, because extractors are imperfect. This holds especially for automated text extraction methods [127]: The best models achieve only 75% in recall on the popular TACRED benchmark [126].
- (4) **Resource limitations:** Human efforts are naturally bounded by available work time, and this applies to a lesser degree also to automated extractors. This is especially relevant in the long tail of knowledge, where social media content emerges continuously, at fast pace.
- (5) **Intractability of materializing all true negations:** For negative assertions, another difficulty joins in: The set of true negative statements is quasi-infinite,³ and thus it is infeasible to materialize negative statements beyond a few salient ones per subject.

We formalize the (in-)completeness of a KB by help of a hypothetical ideal KB K^i . This ideal KB contains all true statements about the domain of interest in the real world. Such an ideal KB is not an easy concept. In the most naive conceptualization, K^i simply contains all statements that are true in reality and that are expressible with a given set of predicates. This makes sense for relations that are sufficiently well-defined such as “sibling” or “place of birth.” However, for other relations (such as “hobby”), this conceptualization may be ill-defined [100]: While one of Albert Einstein’s hobbies was playing the violin, he might have had an unclear number of other “hobbies” (such as

²Making her one of the two people who ever received two Nobel Prizes.

³Infinite if one considers an infinite domain, finite but intractable if one considers a finite domain, e.g., the active domain of a KB.

going for a walk or eating chocolate). Thus, it is not clear what an ideal KB should contain for this relation. The same goes for entities: Is anybody with a doctoral degree a scientist? Or do we count only people employed as scientists? What if a freelancing scientist makes an important discovery? This shows that we can posit the notion of the ideal KB only for well-defined sets of relations and entities: hobbies that are pursued in a registered association, scientists who are employed at a research institution, and so on. While this comes at the expense of what can be targeted, many relations and sets of entities are sufficiently well-defined to establish completeness: all countries that are members of the United Nations; all mountains higher than 1,000 m in a given country; all universities of a given country that can deliver a doctoral degree; and so on. In what follows, we shall focus on such domains where the ideal KB K^i can be reasonably established.

We say that a statement st from the domain of interest is *true in the real world* if $st \in K^i$, and a negative statement $\neg st$ from the domain of interest is true in the real world (“truly false”) if $st \notin K^i$. We say that another KB K is *correct* if $K \subseteq K^i$. Correctness is usually a core focus in KB construction [85] (also referred to as precision, truthfulness, or accuracy). Nevertheless, KBs are of course not all correct at scale. However, since we focus on incompleteness in this survey, we will make the simplifying assumption that the KB at hand is correct.

A selection condition σ is a statement $\langle s, p, o \rangle$, where each component can either be instantiated or a wildcard (“*”). For example, $\langle *, birthPlace, Warsaw \rangle$ is a selection condition. The result of a selection condition σ on a KB K (written $\sigma(K)$) is the set of all statements in K that match σ at its non-wildcard positions. For example, $\langle *, birthPlace, Warsaw \rangle$ selects all people born in Warsaw, and $\langle Marie\ Curie, wonAward, * \rangle$ selects all awards won by Marie Curie.

We can now proceed to define major concepts for our survey.

Definition 2 (KB Completeness). Given a KB K and a selection condition σ , we say that K is *complete* for σ if

$$\sigma(K) = \sigma(K^i).$$

Definition 3 (KB Recall). Given a KB K and a selection condition σ , the *recall* of K for σ is defined as:

$$Recall = \frac{|\sigma(K)|}{|\sigma(K^i)|}.$$

It follows that KB recall is a real-valued concept, while KB completeness is a binary concept, satisfied exactly when recall equals 1.

Terminology. The existing literature does not use terminology consistently. It often uses the terms completeness and recall interchangeably and utilizes others, such as coverage. For this survey, we strictly use *completeness* for the Boolean concept of whether the result of a selection condition on the available KB equals the result of the same selection on the ideal one, and *recall* for its generalization to the quantitative ratio of the two. Formally, a separate notion of completeness is thus superfluous, yet we find the Boolean case frequent enough to indicate it separately.

2.3 World Semantics

In data management, there are traditionally two major paradigms on how to interpret positive KBs: The **closed-world Assumption (CWA)** states that statements that are not in the KB are false in the real world, i.e., $\langle s, p, o \rangle \notin K \Rightarrow \langle s, p, o \rangle \notin K^i$. This contrasts with the **Open-world Assumption (OWA)**, which states that the truth of these statements is *unknown*, i.e., a statement that is not in the KB might or might not be in K^i . The CWA is reasonable in many limited domains, e.g., a corporate database where all employees, orders are known, education management, where

becoming a student requires completing a formal sign-up process, or professional sports where membership in major leagues is well established. In many other settings, however, where it is impossible, unrealistic, or not desired to have all statements and entities of a given domain, the OWA is more appropriate. For example, it is unrealistic, and even undesired, for a KB such as YAGO to contain all people of American nationality. Therefore, most web-scale KBs operate under the OWA.

Example: Consider again *Marie Curie*. Consider the query “Was Marie Curie a Polish citizen?” Under both the OWA and the CWA, the answer would be “Yes,” because the statement $\langle \textit{Marie Curie}, \textit{citizenship}, \textit{Polish} \rangle$ is in the KB. Now, consider the query “Was Mary Curie an Australian citizen?” Under the CWA, given that no such statement is in the KB, the answer would be “no.” Yet, under the OWA, the statement could still be true in reality, so the answer would be “Unknown.” Finally, consider the query for “Nobel Peace Prize winners who are not US citizens.” Under the CWA, that query would return Marie Curie and many others. Under the OWA, this query would have to return the *empty set*, because for any winner, the KB could just be missing US citizenship.

The OWA and the CWA represent extreme stances, and both have severe shortcomings. The OWA makes it impossible to answer “No” to any query, while the CWA produces the answer “No” also in cases where the KB should refrain from taking a stance. Intermediate settings are thus needed, referred to as **Partial-closed World Assumptions (PCWA)** [28, 74], where some parts of the KB are treated under closed-world semantics, others under open-world semantics. One instantiation of the PCWA is the **partial completeness assumption (PCA)** [33, 38]. The PCA asserts that if a subject has *at least one* object for a given predicate, then there are no other objects in the real world beyond those that are in the KB. For example, if the KB knows only one award for *Marie Curie*, then we assume that she won no others. However, if no pet is given, then we assume nothing about her number of pets: She still might have had pets or not. Empirically, the PCA has been found to be frequently satisfied in KBs (see Section 3.2, paragraph “Weak Signals”). Still, the PCA is a generic assumption and does not allow the specification of individual regions of completeness/incompleteness.

2.4 Completeness and Cardinality Metadata

We next introduce two kinds of metadata assertions that can be used to specify areas of the KB that are complete [74, 99].

Definition 4 (Completeness Assertion). Given an available KB K , and an ideal KB K^i , a completeness assertion [74] is a selection condition σ for tuples that are completely recorded. Formally, such an assertion is satisfied if $\sigma(K^i) = \sigma(K)$.

Example: The completeness assertion $\langle \textit{Marie Curie}, \textit{advised}, * \rangle$ specifies that all advisee relations for *Marie Curie* are in the KB. Given the running example from Table 1, this would mean that Curie advised no one other than Marguerite Perey, Óscar Moreno, and the other six listed individuals.

Completeness assertions can be naturally extended to select-project-join queries [74]. For example, one would use a join to specify that the KB knows all the advisees of Physics Nobel Prize winners from the UK. Projections are also possible, e.g., to assert completeness of all subjects that have a VIAF ID, but without recording the ID itself. Both joins and projections come with subtleties, particularly whether subqueries are evaluated on the available KB or on the conceptual ideal resource K^i . Similarly, projections allow to collapse multiplicities of join queries, and completeness

List of Argentine Nobel laureates

From Wikipedia, the free encyclopedia

Argentines have won five Nobel Prizes since 1905. The following is a complete list of Nobel laureates from Argentina:

Laureates [edit]

Year	Name	Field
1936	Carlos Saavedra Lamas	Peace
1947	Bernardo Alberto Houssay Laffont	Physiology or Medicine
1970	Luis Federico Leloir Aguirre	Chemistry

Name	Status	Mapped/checked by
Sandhills, Risinghurst, Barton.		achadwick checked 2009-03-23
Florence Park, Temple Cowley, Lye Valley		achadwick checked 2009-04-06
Cowley Centre South and Oxford Business Park		rjw62
Headington Quarry, Wood Farm		achadwick checked 2009-05-31

Gandhi (1982)

Full Cast & Crew

Directed by [Richard Attenborough](#)

Cast (in credits order) verified as complete

[Ben Kingsley](#) ... [Mahatma Gandhi](#)

[Rohini Hattangadi](#) ... [Kasturba Gandhi](#)

Academic achievements

Up till April 2016, Henrik has authored and co-authored around 350 publications.^[14] Wenzel has been involved in the development and teaching of around 20 graduate and postgraduate courses during his career. He has supervised or co-supervised a total of 15 PhD students. He has managed a total of around 170 research and development projects for a large variety of research organisations, authority agencies, industrial companies and NGOs in more than 10 countries and the EU.

Fig. 1. Examples of completeness and cardinality assertions on the web. Top left and bottom right: Wikipedia. Top right: IMDb. Bottom left: Openstreetmap.⁴

assertions on resulting sets or bags carry different semantics (*Does the KB allow to reconstruct all Nobel Prize winners, or does it also allow to reconstruct how often each won the award?*). We refer the reader to Reference [99] for further detail.

Various human-curated resources allow specifying completeness assertions in the spirit defined above, as shown in Figure 1 (top). On the top left, one can see an assertion about the completeness of Argentinian Nobel laureates. This is an example of a typical join filter: The predicate *wonAward* needs to be joined with the predicate *citizenOf*. On the top right, we see a simpler assertion about cast members of a movie. This corresponds to a selection like $\langle \text{Gandhi}, \text{starsIn}, * \rangle$.

Knowledge bases often give classes special importance (identified via the “type” predicate), and although both can be modeled with the same completeness assertions, one can thus divide completeness further into entity completeness (as in “Do we have all physicists?”) and statement completeness (as in “Do we have all awards for each physicist?”). This distinction has advantages, because the methods to tackle these types of incompleteness are different, as we shall see in Section 3. However, the separation is often not clear-cut, as entities are frequently identified by statements. For example, *French cities* are identified by the selection $\langle *, \text{locatedIn}, \text{France} \rangle$. Therefore, our definition above makes no formal distinction, although works discussed in the following sections may pragmatically focus on one of the two categories of completeness.

Cardinality assertions are inspired by number restrictions from Description Logics [51]. They take the following form:

Definition 5 (Cardinality Assertion). A cardinality assertion is an assertion of the form $|\sigma(K^i)| = n$, where σ is a selection condition, and n is a natural number. It specifies the number of tuples in K^i that satisfy a certain property.

⁴Image sources: <https://wiki.openstreetmap.org/w/index.php?title=Abingdon&oldid=471369>, https://www.imdb.com/title/tt0083987/fullcredits?ref_=tt_ov_st_sm, https://en.wikipedia.org/wiki/List_of_Argentine_Nobel_laureates, and https://en.wikipedia.org/wiki/Henrik_Wenzel

Example: Continuing with the selection condition $\sigma = \langle \text{Marie Curie, wonAward, *} \rangle$, the cardinality assertion $|\sigma(K^i)| = 37$ expresses that, in reality, *Marie Curie* has received a total of 37 awards.

Cardinality assertions can be used to infer recall for a given selection condition by computing the division from Definition 3, $\frac{|\sigma(K)|}{n}$. Moreover, in the special case where $|\sigma(K)| = n$, they can be used to infer completeness.

Example: Assume that we know that, in reality, *Marie Curie* has received a total of 37 awards, i.e., $|\sigma(K^i)| = 37$. In the running example from Table 1, for the selection condition of Marie Curie's awards, $|\sigma(K)| = 2$. Based on Definition 3, this would mean that on her awards, the KB has a recall of $\frac{2}{37} \approx 5.4\%$.

Cardinality assertions are also found in web resources, as shown in Figure 1 (bottom right), where a Wikipedia article mentions the number of PhD students advised by a certain researcher. The snippet at the bottom left contains no explicit numerals but carries a similar spirit of recording to which degree information on districts of a city is recorded in Openstreetmap.

Completeness and cardinality assertions are closely related: Cardinality assertions allow to evaluate completeness, while completeness assertions establish cardinalities. If we know that Marie Curie advised eight PhD students, then we can infer that the KB in Figure 1 is complete. If we know that she held three different citizenships, then we can infer the KB is incomplete. If we know that the KB is complete, then we can infer that the true cardinality of PhD students she advised is seven. Completeness, in turn, enables establishing negation. If we know that the list of Marie Curie's advisees is complete, then we can infer that she did not advise Jean Becquerel, Louis de Broglie, and so on.

Reasoning with completeness metadata. Given complex KB queries and metadata about completeness, a natural question is to deduce whether and which parts of the query answer are implied to be complete (or in the case of negation operations, correct). This problem has received extensive attention in database research [28, 62, 64, 74, 98, 99, 109] as well as knowledge base research [26, 27]. Approaches either take a deductive route (a query result is complete if certain conditions are met) [26, 64, 99] or inductively propagate metadata through query operators [62, 74, 98], with computational complexity mirroring or exceeding the complexity of query answering.

Completeness assertions may also be inferrable via constraints. Most notably, if a predicate is asserted to be functional, then, per subject, presence of one object indicates completeness. For example, if one birth place for Marie Curie is present, then birth places of Marie Curie are complete. Cardinality constraints, like *every person has two parents*, naturally extend this idea and are in turn further extended by the cardinality assertions, which make this subject-specific.

3 PREDICTIVE RECALL ASSESSMENT

In this section, we deal with approaches that can automatically determine the *recall* of a KB, i.e., the proportion of the domain of interest of the real world that the KB covers. We consider the recall of entities (Section 3.1) and the recall of statements (Section 3.2).

3.1 Recall of Entities

To estimate to what degree a KB is “complete,” we would first need to know how many entities from the domain of interest are missing. We can formalize this problem as follows:

Definition 6. Missing Entities Problem

Input: A class C with some instances

Task: Determine how many instances of C are missing compared to K^i .

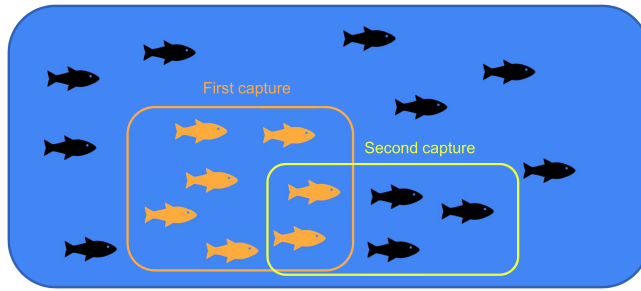


Fig. 2. Mark and recapture: 40% of fish in the second capture are marked \Rightarrow we marked 40% of fish in the first capture \Rightarrow the total number of fish is roughly $7/40\% \approx 17$.

This definition is much less clear-cut than one would hope [84, 100]: Is anyone with a doctoral degree a *scientist*? What is the total number of *cathedrals* in a country if some have been destroyed or rebuilt? What is the total number of *islands* of a country (do we also count *islets* and *rocks*)? What is the total number of inhabitants of a country (do we also count deceased people, do we count only famous people)? Hence, in what follows, we will restrict ourselves to crisp and well-defined classes such as the countries recognized by the UN as of 2022, mountains in a certain country that are taller than 1,000 m, and so on. Let us now look at various methods to address the Missing Entities Problem.

Mark and recapture is a technique for estimating the total estimation of animals in a certain terrain [104]. For example, assume that a biologist wants to know the number N of fish that live in a given pond. For illustration purposes (Figure 2), we assume a small number of fish, with $N = 18$. The biologist captures a number $M = 7$ of fish and marks them (e.g., with a color mark). Then, she releases the fish back into the water (if possible, quickly, to avoid that the number of live fish drops to $N - M = 11$). The next day, she recaptures a number of fish and checks how many of them have marks. Let us say that only 40% of the captured fish are marked. Under the assumption that the ratio $r = 40\%$ of marks in the recaptured population is the same as the ratio of marks in the entire population, the biologist can then estimate the total number of fish by the Petersen estimate as $\hat{N} = M/r$ (which is 17 in our example). This estimator has since been refined by a number of approaches [104]. The question is now how this basic technique can be applied to KBs. It may be tempting to “capture and release” entities from the KB, but the technique works only if the entities are captured from the real world, not from the sample that has already been “captured” by the KB.

Estimators for collaborative KBs [67] extend the notion of mark-and-recapture to collaborative KBs such as Wikidata. The main idea is that, by performing an edit on Wikidata, a contributor samples an entity from the real world. Thus, any statement $\langle s, p, o \rangle$ that a user contributes counts as a sample of s and o from the real world. We can then consider all entities that have been edited during a given time period and that belong to the target class, as one sample. If D is the current number of instances of the target class in the KB, n is the number of entities observed in the sample, and f_1 is the number of entities that have been observed exactly once in the sample, then the Good-Turing estimator estimates the total number of instances of the target class in the real world as $\hat{N} = D \times (1 - f_1/n)^{-1}$. The intuition behind this estimator is as follows: If $f_1 = n$, then every entity in the sample has been seen exactly once. This suggests that, as we keep sampling, we will always see new entities. This means that the size of the real-world population is infinite. If, conversely, $f_1 = 0$, then we have seen all entities at least twice during the sampling, which suggests that the class is not bigger than what we currently have ($\hat{N} = D$). Other possible estimators are

the Jackknife estimator, an estimator with Singleton Outlier Reduction, and an Abundance-based Coverage Estimator [67].

Experiments on the Wikidata KB show that these estimators converge to the ground truth number of missing entities on 5 of the 8 classes that were tested. Interestingly, some of these classes are *composite classes*, i.e., classes defined by a query such as “*Paintings by Van Gogh*.”

Estimators for crowd-sourced KBs [1, 123] target settings where workers are paid to contribute instances to classes or query answers. While every worker has to contribute distinct instances, different workers may contribute the same instances. Workers may also come and go at any moment and have different work styles. If one worker (a “streaker”) adds many entities in one go, then classical recall estimators may overestimate the total population size (mainly because of a high value of f_1 ; see above). One solution to this problem is to cap the number of new entities per contributor to two standard deviations above the average number of entities per contributor.

Experiments with the Crowddb platform⁵ show that the modified estimator is more conservative in its estimation of the total number of entities. Two experiments with a known ground truth were conducted: For “UN-recognized countries,” the new estimator converges faster to the true number. For “U.S. states,” the effect is less visible, most likely because they are easy to enumerate and are thus filled in one go.

Static estimators target KBs where entities are not added dynamically. In such KBs, there is no sampling process from the real world. One way to obtain a lower bound for the true population size is to use Benford’s Law [18]. Benford’s Law says that in many real-world sets of numerical data, the leading digit tends to be small. More precisely, the digit “1” appears in roughly 30% of the cases, the digit “2” in 18% of the cases, and the digit d in $100 \times \log_{10}(1 + d^{-1})$ percent of the cases. This applies in particular to quantities that grow exponentially (such as the population of a city), because a quantity to which a multiplicative factor is applied repeatedly will run through the leading digit “1” more often than through the other digits. To apply this technique to KBs [114], the target class has to have a numerical predicate, e.g., the population size for cities, the length for rivers, the price for products, and so on. This predicate has to obey certain statistical properties for Benford’s Law to work. We can then collect the first digits of all numbers and check whether the distribution conforms to Benford’s Law. In the work of Reference [114], this is done primarily to quantify to what degree the sample of entities in the KB is representative of the real-world distribution. However, the data can also be used to compute the minimum number of entities that we would have to add to the KB for the distribution to conform to Benford’s Law. Under the assumption that the real-world distribution follows Benford’s Law, this number constitutes a lower bound for the number of missing entities.

Experiments show that a parameterized version of Benford’s Law applies to a number of very diverse predicates, including the population of places, the elevation and area of places, the length and discharge of water streams, the number of deaths and injured people for catastrophes, and the out-degree of Wikipedia pages.

3.2 Recall of Statements

After having discussed the recall of entities, let us now turn to the completeness of statements. More precisely:

Definition 7. Missing Object Problem

Input: A knowledge base K , a subject s , and a predicate p

Task: Determine if there is o with $\langle s, p, o \rangle \in K^i$ and $\langle s, p, o \rangle \notin K$.

⁵<http://www.crowddb.org/> down as of September 9, 2022.

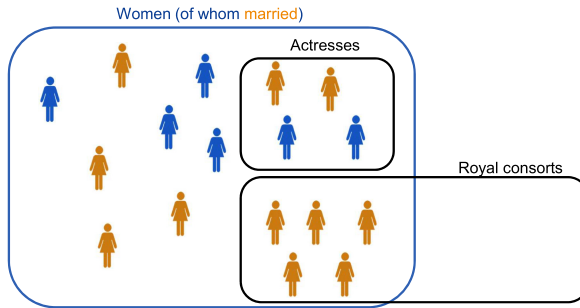


Fig. 3. To determine whether being married is an obligatory attribute of the class *Woman*, we check whether its distribution in the KB changes when we walk into subclasses.

As an example, consider *Marie Curie* and the predicate *wonAward*. We are interested in finding whether she won more awards than those given in the KB. It is not always easy to define what we consider missing objects [100]: Does an award from her high-school count? Is a public recognition tantamount to an award? Is a rejected award still an award she won? and so on. In what follows, we assume that we can determine whether a relationship holds, e.g., by a vote from crowd-workers.

A first hint on missing objects comes from **obligatory attributes**. An obligatory attribute of a given class is a predicate that all instances of the class must have in the real world at the current point in time. For example, *birthDate* is an obligatory attribute for humans, while *hasSpouse* is not. If we knew the obligatory attributes, then we could use them to solve parts of the Missing Object Problem: If an instance of a class has no object for an obligatory attribute, then it is necessarily incomplete on that attribute. It turns out that the obligatory attributes of a class can be determined automatically from the incomplete KB [61]. For this purpose, we look at the ratio of instances of the class that have the predicate in the KB. For example, we can look at the ratio of women who are married (Figure 3). We then check if this ratio changes when we go into subclasses of the target class or into intersections of the target class with other classes. In our example, we can go into the subclass of *Actresses* or into the intersecting class of *Royal Consorts*. If the ratio changes, then a theorem tells us (under a number of assumptions) that the predicate *cannot* be obligatory. In our example, the ratio of married women changes when we go into the subclass of *Royal Consorts*. Hence, *married* cannot be an obligatory attribute. If we assume that the other predicates are all obligatory, then we can spot places in the KB where objects must be missing.

Experiments on YAGO and Wikidata show that the approach can achieve a precision of 80% at a recall of 40% over all predicates. The approach can also determine that, for people born before a certain time point, a *deathDate* becomes an “obligatory” attribute.

Weak signals can be combined to estimate whether a given subject and a given predicate are complete in a given KB [36]. The simplest of these signals is the **Partial Completeness Assumption (PCA)**, which says that if the subject has some objects with the predicate, then no other objects are missing. Other signals can come from the dynamics of the KB: The *No-Change Assumption* says that if the number of objects has not changed over recent editions of the KB, then it has “converged” to the true number, and no more objects are missing. The *Popularity Assumption* says that if the subject is popular (i.e., has a Wikipedia page that is longer than a given threshold), then no objects are missing. A more complex signal is the *Class Pattern Oracle*. It assumes that if the subject is an instance of a certain class *c*, then there are no missing objects. For example, if the subject is an instance of the class *LivingPeople*, then no death date is missing. The *Star Pattern Oracle* does the same for predicates: If, e.g., a person has a death place, then the person should have a death date.

These signals can be combined as follows: We first add the simple signals as statements to the KB. For example, if *Elvis Presley* is known to be popular, then we add $\langle \textit{ElvisPresley}, \textit{is}, \textit{popular} \rangle$ to the KB. If Elvis has won several awards, then we add $\langle \textit{ElvisPresley}, \textit{moreThan}_k, \textit{wonAward} \rangle$ for small k . Then, we add the ground truth for some of the subjects. For example, if we know that *Elvis Presley* has won no more awards than those in the KB, then we add $\langle \textit{ElvisPresley}, \textit{complete}, \textit{wonAward} \rangle$ to the KB. Finally, we can use a rule-mining system (such as AMIE [38]) to mine rules that predict this ground truth. Such a system can find, e.g., that popular people are usually complete on their awards, as in $\langle s, \textit{is}, \textit{popular} \rangle \Rightarrow \langle s, \textit{complete}, \textit{award} \rangle$. These rules can then be used to predict completeness for other subjects, i.e., to predict whether Marie Curie has won more awards than those mentioned in the KB.

Experiments with a crowd-sourced ground truth on YAGO and Wikidata show that some predicates are trivial to predict. These are obligatory attributes with only one or few objects per subject: *bornIn*, *hasNationality*, *gender*, and so on. Here, the CWA and PCA work great. For the other relationships, the Popularity Assumption has a high precision, but low recall, i.e., misses that many more subjects than predicted are complete. The Star- and Class-Oracles also do well. The combined approach can achieve F1-values of 90%–100% for all 10 predicates that were tested, with the exception of *hasSpouse*: It remains hard to predict whether someone has more spouses than are in the KB.

Textual information can also be used to spot incomplete areas of the KB. For example, assume that we encounter the sentence “Marie Curie brought her daughters Irène and Eve to school.” Then, we can conclude that Curie had at least two daughters. In common discourse, we would even assume that she had no other daughters and in fact no other children at all (if she had other children, then we would expect the speaker to convey that, e.g., by saying “brought only her daughters to school”). This conclusion is an *implicature*, i.e., an information that is conveyed by an utterance (or text) even though it is not literally expressed [45]. In what follows, let us consider a given sentence about a given subject and a given predicate, and let us assume a KB that is complete for that subject and predicate (e.g., Wikidata for popular subjects). If we use a simple open information extraction approach, then we can extract the objects that the sentence mentions for the predicate and compare them to the objects that the complete KB contains [97]. If the sentence mentions all objects, then we consider it *complete*. In this way, we can build a training set of complete and incomplete sentences for a given predicate, and we can train a classifier on this set. This classifier can then be used to determine if some other sentence is complete, and if it is, then we can determine if the subject of that sentence is complete in the KB.

Experiments across 5 predicates show that the approach works better on paragraphs than on sentences, because in many cases, the objects are enumerated across several sentences. The F1-values are 45%–75%.

3.3 Summary

The recall of a KB can be estimated for entities and for statements. Different approaches have been developed for both cases (Table 2) with promising results. However, the approaches for entities suffer from a small ground truth: In only few cases, the total number of instances of a class is known. This makes the experiments difficult to judge.

4 CARDINALITIES FROM KBS AND TEXT

We move from counting entities and objects to identifying cardinality assertions. As defined in Section 2, cardinality assertions specify the number of records in the ideal KB K^i that satisfy a certain property. But K^i is a hypothetically ideal KB, making the computation of cardinality assertions through selection conditions non-trivial.

Table 2. Overview of Predictive Recall Assessment Approaches

Approach	Target	Assumptions
Mark & Recapture [104]	Entities	Sampling from the real world
Collaborative KB Estimator [67]	Entities of a given class	Each fact is a random draw from the real world
Crowd-sourced KB Estimator [123]	Entities of a given class	Crowd-workers contribute entities independently
Static Estimator [18]	Entities of a given class	Distribution follows Benford's Law
Obligatory Attribute Estimator [61]	Relations that are obligatory in the real world	Facts are sampled i.i.d. from the real world
Weak Signal Estimator [36]	Missing objects for a given subject and relation	Sufficient number and quality of weak signals
Extraction from Text [97]	Number of objects for a given subject and relation	Explicit mentions of objects or numbers in text

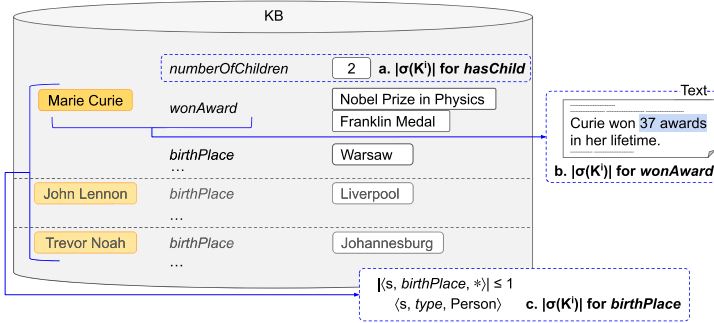


Fig. 4. Illustrating three ways of obtaining cardinality assertions (a) from KB predicates (for the number of children), (b) from text (for the number of awards), and (c) from inference over KB statements (for the number of birthplaces).

Cardinality assertions are expressed explicitly through statements in KB, $\langle \text{Marie Curie}, \text{numberOfChildren}, 2 \rangle$, and in text, "Curie won 37 awards in her lifetime." Correctly identified, such statements allow for a direct numeric evaluation of KB recall. Alternatively, if we know that a certain predicate is complete for a certain subject, then we can use that knowledge to deduce the cardinality assertion. In this section, we tackle the problem of identifying cardinality assertions from the KB and text, illustrated in Figure 4. We look into methods that identify predicates that store cardinality assertions explicitly, for instance, $\langle s, \text{numberOfChildren}, n \rangle$, and statements that ground the cardinality into objects, such as $\langle s, \text{hasChild}, o_i \rangle$, $i \in [0, n]$. From now on, we call the predicates that store cardinality explicitly *counting predicates*, p_c , and the predicates that store cardinality grounding objects *enumerating predicates*, p_e .

4.1 Cardinality Information in KBs

The most obvious way to obtain cardinality assertions from KBs would be to use aggregate queries. For example, to determine the number of children Marie Curie had, we can count the number of statements satisfying the selection condition, $\langle \text{Marie Curie}, \text{wonAward}, * \rangle$, but, as we have seen in the previous section, KBs suffer from the Missing Objects Problem. Fortunately, some KBs express cardinality assertions explicitly through counting predicates. For example, we find the predicates *numberOfGoals* (for the goals scored by football players) and *totalHurricanes* (for the number of hurricanes in a cyclone) in DBpedia and *numberOfEpisodes* (for a film/TV/radio series) in Wikidata as well as in YAGO. Unfortunately, the predicates of these cardinality assertions usually do not follow any naming scheme: Some predicates have the prefix *numberOf...*, but others are specific to the class at hand, including *staffSize*, *population*, or *member count*. Furthermore, KBs are normally unaware of the semantic relation between the enumerating predicate *hasChild* and the counting predicate *numberOfChildren*, and they do not coexist for all entities.

Usually, a counting predicate (such as *numberOfChildren*) links a subject to the number of objects of the corresponding enumerating predicate (*hasChild* in this case). However, a counting predicate may also concern the number of *subjects* of some enumerating predicate. For example, the enumerating predicate *worksAt* links a person to their workplace, but the counting predicate *numberOfEmployees* links the workplace to the number of people who work there. To deal with such cases, we assume that the KB contains, for each predicate p , also its inverse p^{-1} , with all triples. In our example, we assume that the KB contains also the predicate *worksAt*⁻¹ (which we call *hasEmployee*). Then, *numberOfEmployees* is the counting predicate of the enumerating predicate *hasEmployee*.

Our first task here is to identify these two classes of predicates. Our second task is then to identify a mapping between the two sets of predicates.

Definition 8. Cardinality Detection Problem

Input: A knowledge base K

Task: Determine the counting predicates p_c and the enumerating predicates p_e . Also determine which p_e corresponds to which p_c .

Cardinality bounding. A first step for cardinality detection can be to bound the cardinality of the predicates. For example, if we know that there are 200 countries in our knowledge base, then we know that a person cannot have more than 200 citizenships. This knowledge can then help us find counting predicates. In some cases, cardinality bounds come from the KB itself. For instance, Wikidata uses the *single-value constraint*⁶ for predicates with exactly one object.⁷ However, such cases are rare.

If the KB can also contain incorrect entries (more realistic than the naive correctness assumption that was brought forward in Chapter 2), then one approach to automatically bound cardinalities is by mining significant maximum cardinalities of a predicate for a given class [43]. For a given predicate, *parent*, if we see that a significant portion of the entities in the class *Person* have two *parent* objects, then we can say that the predicate has a maximum cardinality of 2. This cascades to the subclasses of *Person*, for instance, to scientists, physicists, and so on, unless any of the subclass has a tighter bound for significant number of its members. This is closely related to outlier detection, but with a focus on mining generic constraints, with using them for outlier detection being only one of several possible applications. For example, with the above constraint in mind, entities of type person with six parents (which are relatively few) could be ignored. Using Hoeffding's inequality constraint [48], a significant maximum cardinality of a class can be reliably mined for a given confidence level and a minimum likelihood threshold. The approach works top down, starting from the class going deeper into its subclasses and pruning very specific constraints for which Hoeffding's inequality would no longer hold true. This approach works well for finding maximum cardinalities for functional predicates; for instance, all humans have one birth year or all football matches have two teams, since these are more stable across entities of a class: Majority of persons have one or two parents and very few have more than two. We will see how to tackle entity-variant cardinalities, such as books by an author or destinations of an airline, which can vary between entities, even those belonging to the same class in the next approach.

Cardinality predicate detection aims at identifying existing cardinality assertions and the corresponding enumerating predicates in the KB. A naive approach would be to identify all predicates with positive integers objects as *counting predicates* and all predicates with one or more

⁶Single-valued constraint definition in Wikidata: https://www.wikidata.org/wiki/Help:Property_constraints_portal/Single_value

⁷An element can have only one atomic number <https://www.wikidata.org/wiki/Property:P1086>

KB entities as *enumerating predicates*. This naive approach fails due to a number of reasons. First, a positive integer value is a necessary but not sufficient condition for counting predicates. For instance, KB predicates that store identifiers (*episode number*, *VIAF ID* in Wikidata), measurement quantities (*riverLength*), or counts of non-entities (*floorCount*) are not counting predicates. Second, functional predicates such as *birthPlace*, *atomicNumber*, and so on, do not commonly occur with counts. While these predicates do enumerate fixed objects, they do not have any meaningful counting predicate: The fact that the number of birth places of *Marie Curie* is one is not informative. Third, many quasi-functional predicates are predominantly functional but also take multiple values, for instance, citizenship is primarily single-valued with many famous exceptions.⁸

Signals such as the predicate domain, the predicate range, and textual information can be utilized to identify the cardinality predicates [42]. Predicate names provide some clues as to whether it might be an enumerating predicate. For instance, the word *award* in its plural form occurs almost as frequently, if not more than the word in its singular form, but this is not true for birth place (birth places). Objects of enumerating predicates are also entities, and encoding the type information of the subjects and the objects of a predicate can also provide clues: A type *Person* (*Curie*) usually co-occurs with multiple instances of the type *Award* (*Nobel Prize in Physics*, *Franklin Medal*). Range statistics such as the mean and the percentile values are also important clues: In DBpedia, the number of statements per subject for the predicate *wonAward* is 2.8, on average, and the mean and the 10th percentile values of the predicate *doctoral students* is 28.3 and 5.5, respectively.

Predicate alignment aims to align enumerating predicates with corresponding counting predicates. This would allow us to match cardinality assertions with the statements grounding the cardinality. Aligned predicates can be used to estimate the KB recall, since counting predicates define expected (or ideal) counts for enumerating predicates. If we know that *award* aligns with *numberOfAwards*, then for all entities with *numberOfAwards*, we can compute the recall of their *award* statements. In reality, exact matches such as the one above exist only for very few entities and predicates. More often than not, the enumerations are incomplete and overlapping. For instance, an institute may use *numberOfStaff* or *numberOfEmployees* to mean the same thing, and the corresponding enumerations could come from $workInstitution^{-1}$ and $employedBy^{-1}$. Heuristics such as exact and approximate co-occurrence metrics and linguistic similarity of the predicate labels can be used to suitably aligned pairs [42]. The number of alignments in a KB thus obtained is much lower than the number of counting and enumerating predicates. For instance, there is no counting predicate that aligns with the enumerating predicate *wonAward*. These alignments are directional as well. For instance, if the score of $(workInstitution^{-1}, academicStaff)$ is greater than that of $(academicStaff, workInstitution^{-1})$, then an entity with the enumerating predicate $workInstitution^{-1}$ is more likely to have the counting predicate *academicStaff* than the other way around.

4.2 Cardinality Information from Text

So far, we have looked into KBs for cardinality detection, but the KBs are usually sparse. Even if we have identified that *numberOfDoctoralStudents* is a counting predicate, we cannot predict its value for an entity that has the enumerating predicate *doctoralStudent* but no counting predicate *numberOfDoctoralStudents*. In such cases, we can turn to textual data, i.e., we can tap into textual sources for retrieving cardinality information. For example, if a text says “*Marie Curie advised 7 students*,” then we want to extract the number 7 for the predicate *advised*. More formally:

⁸Scientists with more than one citizenship during their whole life <https://w.wiki/5UR3>

Definition 9. Counting Quantifier Extraction Problem**Input:** A subject s , a text about s , and a predicate p **Task:** Determine the number of objects with which s stands in predicate p from the text.

The first challenge we face is the *variability of text*: The same number can be expressed either as a numeral (“3”), as a word (“three”), or as an expression (“trilogy of books”). The second challenge is *compositionality*: Quantities may have to be added up, as in “... *authored three books and 20 articles*,” which indicates that the predicate *numberOfPublications* has to have the value $3+20=23$. Finally, we have to *determine the target predicate*: The sentences “*advised 7 students*” and “*supervised 7 students*” both map to the target enumerating predicate *advised*. Several approaches can be used to overcome these challenges.

Open IE methods [25, 69] aim to extract relational statements from text, and these can also include cardinality information. For example, from the sentence *Marie Curie has two children*, these methods can extract $\langle \text{Marie Curie, has, two children} \rangle$. However, this information is not linked to the KB predicate *hasChild*, and it is also not trivial to recover the number 2. Roy et al. [106] take this forward by proposing an IE system that targets quantities. It extracts standardized quantity statements of the form $\langle \text{value, units, change} \rangle$. This approach works well and can in principle be applied to cardinality assertions. The **never-ending learning paradigm (NELL)** [21] has seed predicates that capture predicate cardinality, such as the *numberOfInjuredCasualties*, but, since these are tied to specific cardinality values, it does not learn them for future extractions.

Current state-of-the-art NLP pipelines have made cardinality detection in texts much easier.⁹ Even though Open IE extractions contain cardinality assertions, extracting the correct count value and mapping to KB predicates is an open challenge. Next, we discuss several approaches that target specifically the extraction of cardinalities.

Manually designed patterns can be used to extract cardinality assertions from text [72]. The idea is to capture compositionality and variability of cardinal information in text that is outside the capability of current Open IE methods. Reference [72] proposes 30 manually-curated regular expression patterns for the *hasChild* predicate in Wikidata the extractions achieve more than 90% precision when compared with a small set of manually-labelled gold dataset and a larger KB available silver dataset. Cardinality assertions were extracted for 86.2k humans in Wikidata, of which only 2.65% had complete *hasChild* statements.

Automatic extraction of cardinalities can be modeled as a sequence labelling task where each predicate has its own model. The work of Reference [71] explores one feature-based and one neural-based conditional random field model. The cardinal information in the input sentence is replaced with placeholders, namely, *cardinal*, *ordinal*, and *numeric term* so the model does not learn specific tokens but the concept. The models then input the features such as n-grams and lemmas in the context-window of the placeholder to the feature-based CRF model and in case of a bi-directional LSTM model the input comprises the words, placeholders, and character embeddings of the tokens. The model outputs the labels *COUNT* for cardinals, *COMP* for composition tokens, and *O* for other tokens, with a confidence score. A sentence can have multiple cardinal labels or multiple sentences can have cardinal information for a given s and p . Hence, a consolidation step aggregates cardinals around compositional cues. A cardinality assertion can occur multiple times in a text, in which case, a heuristic ordering is applied such that cardinal > numeric term > ordinal > article, breaking ties with confidence scores. This follows the intuition that if a text has a sentence “*Curie published 15 articles.*” and another sentence “*Curie published her first article in 1890.*”, then 15 would be the predicted cardinality for Curie’s publications. Existing

⁹<https://spacy.io/usage/linguistic-features>

cardinal predicates in the KB can be selectively used as ground truth to evaluate the precision of the automatically extracted cardinalities. For example, one can restrict the KB to popular entities, and one can exclude cardinalities above the 90th percentile value, as these are most likely outliers. One must be selective, since incomplete KB information can negatively impact overall training and evaluation.

Given that LLMs have transformed a range of tasks, an obvious question is whether they can be useful for the extraction of cardinality assertions. This broad question comes in two variants: (i) Can LLMs help to extract cardinality assertions from a given text? (ii) Based on the text they have seen during pre-training, can LLMs directly output cardinality assertions?

The answer to (i) appears a confident yes, as several studies have shown that LLMs can be used as powerful components in textual information extraction [5, 24, 57]. Yet, such a usage still requires comprehensive efforts around data selection and preparation, information retrieval, and output consolidation, and thus, does not fundamentally transform the task.

A fundamental transformation could be achieved in the second paradigm. As LLMs have seen huge amounts of text during pre-training, one can try to extract relations directly from these models without further textual input at the extraction stage. Although early works have found mixed results [65], current LLMs like GPT-4 can quite reliably assert counts in the commonsense domain, e.g., counts of body parts (feet, wings, ...) of different animals. For extracting counts of named entities, similar observations hold: As long as the counts are frequently asserted in text, LLMs can return them. However, once the LLMs would need to aggregate themselves, they fall back to common numbers in the relation of interest [113]. Although LLMs are becoming incrementally more powerful on benchmarks, it appears that Transformer-based architectures exhibit principled limitations, which makes correctly solving count tasks over text difficult [46].

4.3 Summary

In this section, we have defined the task of identifying cardinalities for predicates. We have seen how to extract cardinalities from KBs and from textual data. Cardinality information has many applications and yet unsolved challenges. We have compiled the related works covered in this section, their focus, strengths, and limitations in Table 3.

Applications. The CounQER system [41] demonstrates the usefulness of **aligned predicates** in a simple QA setting where, given a subject s and (an enumerating) a counting predicate p , the system returns the objects that satisfy $\langle s, p, * \rangle$ and statements from top-five aligned (counting) enumerating predicates if available. For instance, in DBpedia, if we look for the *Royal Swedish Academy of Sciences*, then the enumerating predicate *workplaces*⁻¹ returns eight entities who work there, but we also learn that the aligned counting predicates *academicStaff* and *administrativeStaff* are unpopulated. Again in DBpedia, we learn that the enumerating predicate *doctoralStudent* has no corresponding counting predicate: For the subject *Marie Curie* and the predicate *doctoralStudent*, the system returns seven entities who were her doctoral students but no aligned cardinality assertions.

Tanon et al. [86] use cardinality information to improve rule-mining. Rule mining is the task of finding interesting associations between entities in a KB, such as: If $\langle ?x, hasSpouse, ?y \rangle$ and $\langle ?y, livesIn, ?z \rangle$, then $\langle ?x, livesIn, ?z \rangle$. Such rules can then be used to predict the place of residence of someone. Cardinality information can avoid that we predict too many such places of residence per person by downgrading the scores of predictions that violate (soft) cardinality constraints. The authors evaluate the recall-aware scores against standard scoring metrics and find that recall-aware scores highly correlate with rule quality scores in the setting of increasingly complete KBs. Another line of work uses cardinality information as priors in neural link prediction [75]. Similar

Table 3. Comparison of Related Works on Cardinality Information

Source	Work	Focus	Strengths	Limitations
KBs	Giacometti et al. [43]	Extract maximum cardinalities for class and predicate pairs.	Provides significance guarantees. Scalable for large KBs. Efficient pruning to reduce search space.	Majority predicates identified have maximum cardinality of 1. Scope for discovering non-functional assertions is limited.
	Ghosh et al. [42]	Extract counting and enumerating predicates.	Provides important features for identifying cardinality predicates. Maps counting predicates to matching enumerating predicates.	Statistical cues provide weak signals. Textual cues have limited informativeness.
Open IE	Mausam et al. [69] Corro et al. [25]	Extract all $\langle s, p, o \rangle$ statements.	Captures cardinality information present as counts.	Canonicalization required to incorporate statements in to KBs.
	Carlson et al. [21]	Extract new statements and rules.	Never-ending learning paradigm (NELL). Infers new predicates.	Cannot learn cardinality assertions from seed predicates.
	Roy et al. [106]	Extract quantity statements.	In principle, can extract cardinality assertions.	Has only been used for quantity statements.
Text	Mirza et al. [72]	Compute cardinality assertions.	Manual patterns are effective in cardinality extractions.	Scaling predicate-specific patterns.
	Mirza et al. [71]		Sequence modeling for automatic extraction of cardinality information. Consolidation across compositions and multiple mentions.	Prior knowledge of cardinality predicates. Zero counts extraction is limited to ad hoc preprocessing.

to the work by Tanon et al. [86], they regularize the number of high-probability predictions by penalizing the model when the number of predictions violate the cardinality bounds of a given relation type.

Open challenges. LLMs could in principle be used to infer commonsense cardinality information such as the number of parents a person has or specific cardinality assertions, such as Marie Curie advised seven doctoral students. Experiments in probing older LLMs for numerical commonsense [65] show that fine-tuning improves the model performance, though the models could not surpass humans. The 2023 edition of the LM-KBC challenge [111], whose main focus is on constructing KB from an LLM, contains cardinality prediction for two relations, *numberOfChildren* of a person and *numberOfEpisodes* of a TV series. Even the best-performing system, which relied on GPT-4, achieved only 69% F1-score on both relations. It appears that more work is needed here before LLM outputs could reliably feed a knowledge base. A common challenge to both tasks of extracting cardinality assertions and aligning counting with enumerating predicates is generating high-quality training and ground truth data. There is no single authoritative source on groundtruths: We have IMDB for movies,¹⁰ cast, and crew and the GeoNames dataset for geographical locations.¹¹ These are great examples of high-recall datasets, but the situation is not as rosy when we move to other topics, such as scientists or monuments. Crowd-sourcing is an option, but it poses several challenges, most importantly data quality and scalability. The other option commonly used is to employ heuristics. As we saw for cardinality extraction from text [71], distant supervision can be used to extract ground truth statements, with certain restrictions, such as relying on popular entities and KB statistics, such as the 90th percentile predicate value to filter out possible outliers.

¹⁰IMDB: <https://www.imdb.com/>

¹¹GeoNames <https://www.geonames.org/>

5 IDENTIFYING SALIENT NEGATIONS

Knowledge bases store by and large positive knowledge and very little to no negative statements. This happens for principled reasons, as the set of possible negations is vast and possibly infinite (depending on whether one assumes a finite or infinite set of constants). Adding complete sets of negative statements is therefore hardly a goal. Many standard AI applications, such as question answering and dialogue systems, would often benefit from statements about popular entities, i.e., explicit negations for salient cases. For instance, a cooking chatbot should be aware that certain ethnic food is not meant to be heated, e.g., *Hummus* and *Gazpacho*, and a general-purpose search engine must be confident about common factual mistakes, such as famous people not winning certain awards in their domains, e.g., *Stephen Hawking* and *The Nobel Prize in Physics*. Completeness statements enable the inference of negations, yet are themselves hard to come by. Furthermore, even though negative statements are in principle much more numerous than positive statements, only few of them are interesting. In this section, we approach the problem of negation materialization, therefore, not with the goal of completeness, but with the goal of a *high recall among salient negations*.

The concept of *salience* has a long history in psychology [121] as well as in sociolinguistics [91], and there are also recent attempts to capture it in knowledge bases [59]. Approaches to modeling salience typically revolve around the concepts of frequency, unexpectedness, or self-interest (for acting agents), yet universal agreement is lacking. Furthermore, none of the models is easy to operationalize. The works that we present next, therefore, usually utilize human (crowd-worker) judgments as yardstick for salience.

We first review existing negative knowledge in open-world knowledge bases. We then show how *salient negative statements* can be automatically collected from within incomplete KBs and via text extraction.

5.1 Negation in Existing KBs

Web-scale KBs operate under the **Open-world Assumption (OWA)**. This means that an absent statement is not false but only unknown. The only ways to specify negative information is to either explicitly materialize negative statements in the KB or to assert constraints that implicitly entail negative statements. In this section, we focus on the former case. Even though most KB construction projects do not actively collect negative statements, a few of them allow implicit or explicit negative information:

- Negated predicates: A few KBs contain predicates that express negative meaning, i.e., contain negation keywords. For example, DBpedia [13] has predicates such as *carrierNeverAvailable* for phones and *neverExceedAltitude* for airplanes. The medical KB KnowLife [34] contains predicates such as *isNotCausedBy* and *isNotHealedBy*. Wikidata allows a few type-agnostic negated predicates, namely, *differentFrom* (827000 statements), *doesNotHavePart* (535 statements), *doesNotHaveQuality* (422 statements), *doesNotHaveEffect* (36 statements), and *doesNotHaveCause* (13 statements). A more systematic example for negated predicates can be found in ConceptNet [116], where the 6 main predicates have negated counterparts, namely, *NotIsA*, *NotCapableOf*, *NotDesires*, *NotHasA*, *NotHasProperty*, and *NotMadeOf*. Yet, the portion of negative knowledge is less than 2%. Furthermore, many of the negative statements are uninformative, as in $\langle \textit{tree}, \textit{NotCapableOf}, \textit{walk} \rangle$.
- Count predicates: A subtle way to express negative information is by matching count with enumeration predicates (see Section 4.1). For example, if a KB asserts $\langle \textit{Marie Curie}, \textit{numberOfChildren}, 2 \rangle$ accompanied by two *hasChild*-statements, then this indicates that for this subject-predicate pair, the list of objects is complete. Therefore, no other entity is a child of *Curie*.

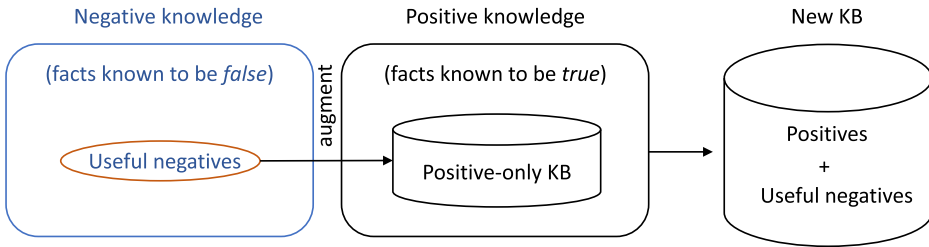


Fig. 5. The main challenge is to identify the subset of *useful* negative statements that can be added to an existing positive-only open-world KB.

- Statements with negative polarity: In the Quasimodo KB [105], every statement is extended by a polarity value to express whether it is a positive or a negative statement, e.g., $\langle \text{scientist}, \text{has}, \text{academic degree} \rangle$ with $\text{polarity}=\text{positive}$. Quasimodo contains a total of 351k negative statements.
- No-value objects: Wikidata [125] allows the expression of universally negative statements, where a subject-predicate pair has an empty object. For example, $\langle \text{Angela Merkel}, \text{hasChild}, \text{no-value} \rangle$.¹² The total number of such statements with a no-value object in Wikidata is 20.6k
- Deprecated rank: KBs like Wikidata encourage flagging certain statements as incorrect as opposed to removing them. These are usually outdated statements or statements that are known to be false,¹³ with a total of 20.4k statements.

While these notions give us a way to express negative statements, one still has to find a way to identify salient ones, such as a well-known physicist not winning the *Nobel Prize*, namely $\neg \langle \text{Stephen Hawking}, \text{wonAward}, \text{Nobel Prize in Physics} \rangle$. A key challenge is that the number of false facts or statements is much larger than the one of positive facts, e.g., Stephen Hawking studied at 4 educational institution, versus thousands that he did not study at. An illustration of the research problem is shown in Figure 5. Among the large set of negative knowledge, the key is to identify the subset of *useful* negatives, and add them to the existing positive-only KB.

As discussed above, we take a model-free approach here and leave the choice of what is salient to human annotators. We hypothesize only that frequency and unexpectedness are likely ingredients [91, 121] of their decision. Where human judgments cannot be obtained, one could resort to model-based metrics of unexpectedness that can be computed automatically, like proposed in [59], though at the risk of optimizing for what can be conveniently computed, instead of what is truly salient.

Definition 10. Salient Negations Problem

Input: A subject s in a web-scale incomplete KB

Task: Identify accurate and salient negative statements about s .

We divide approaches into three main categories: methods that use well-canonicalized KBs, methods that use loosely structured KBs, and methods that use text corpora as the main source of negative statements.

5.2 Salient Negations in Well-structured KBs

Famous KBs such as Wikidata, YAGO, and DBpedia consist of well-canonicalized statements (with minimal ambiguity), i.e., the A-box, and are accompanied with manually crafted schemas (a.k.a. the T-box). The following approaches discover interesting negations about entities by relying on positive statements from well-structured KBs:

¹²<https://www.wikidata.org/wiki/Q567>

¹³<https://www.wikidata.org/wiki/Help:Deprecation>

Peer-based inferences [8] is one of the earliest approaches to solve this problem. It proposes deriving candidate salient negations about a KB entity from *highly related entities*, i.e., peers, then ranks them using relative statistical frequency. For example, *Stephen Hawking*, the famous physicist, has never won a Nobel Prize in Physics, nor has he received an Academy Award. However, highly related entities (other physicists who are Nobel Prize winners) suggest that not winning the Nobel Prize is more noteworthy for Hawking than not winning the Academy Award. The first step must thus be to identify the relevant peers for the input entity. The method of Reference [8] offers three similarity functions to collect peer entities for a given input entity: (i) class-based similarity [15]: This takes advantage of the type-system of the KB by considering two entities as peers if they share at least one type; (ii) graph-based similarity: This relies on the number of predicate-object pairs that two entities have in common; and (iii) embedding-based similarity: This captures latent similarity between entities by measuring the cosine similarity of their embeddings [129].

Example: Using (i), *Stephen Hawking* is a physicist like *Max Planck* and *Albert Einstein*, hence, information about *Planck* and *Einstein* can help in creating candidate negations for *Hawking*. Using (ii), *Stephen Hawking* and *Boris Johnson* share 9 predicate-object pairs, including (*gender, male*), (*citizenship, U.K.*), and (*native language, English*). Using (iii), one of the closest entities to *Hawking* using this measure is his daughter *Lucy Hawking*.

Negations at this point are only candidates. Due to the KB's incompleteness, each candidate could be true in the real world and just be missing from the KB. The peer-based method requires the candidates statements to satisfy the partial completeness assumption (the subject has at least one other object for that property [37, 38]) to be considered for the final set. In particular, if *some* awards are listed for the subject *Stephen Hawking*, then the list of awards is assumed to be *complete*, and any missing award is absent due to its *falseness*.

Example: We know that *Hawking* has won other awards such as the *Oskar Klein Medal* and has children including *Lucy Hawking*, but we know nothing about his *hobbies*. Therefore, the candidate statement $\neg\langle\textit{Stephen Hawking}, \textit{hasHobby}, \textit{reading}\rangle$ is discarded.

In an evaluation over correctness of inferred negations [8], this simple yet powerful rule increases the accuracy of results by 27%. Remaining candidate negations are finally scored by relative peer frequency. For instance, 2 out of 2 peers of *Hawking* have won the *Nobel Prize in Physics* but only 1 out of 2 is the parent of *Eduard Einstein*.

Order-oriented peer-based inferences [10] is an extension of the previous method where KB qualifiers such as temporal statements are leveraged to obtain better peer entities and provide negations with explanations. In this order-oriented method, one input entity can receive *multiple* peer groups. For instance, three separate peer groups for *Max Planck* are *winners of the Nobel Prize in Physics*, *winners of Copley Medal*, and *alumni of Ludwig Maximilian University of Munich*. In addition to peer frequency, the ranking also accounts for the proximity between the input and peer entities.

Example: The negative statement $\neg\langle\textit{Max Planck}, \textit{educatedAt}, \textit{The University of Cambridge}\rangle$ with provenance “unlike *the previous 3 out of 3 winners of the Nobel Prize in Physics*” is favored over $\neg\langle\textit{Max Planck}, \textit{citizenOf}, \textit{France}\rangle$ with provenance “unlike *3 out of the previous 18 winners of the Nobel Prize in Physics*.”

In this example, *temporal recency* is rewarded. In other words, the same peer frequency that is far behind in the ordered peers will receive a lower score. Moreover, this work introduces the notion of *conditional negation*, as opposed to the previous simple ones. While a simple negation is expressed using 1 negative statement, $\neg\langle\textit{Albert Einstein}, \textit{educatedAt}, \textit{Harvard}\rangle$, a conditional negation goes beyond 1 to express negative information that is true *only* under certain condition(s).

Example: $\neg\exists o \langle \text{Albert Einstein, educatedAt, } o \rangle \langle o, \text{ locatedIn, U.S.} \rangle$ (meaning *Einstein* has never studied at a university in the *U.S.*).

In a crowdsourcing task, to evaluate the quality of result negations in Reference [10], the peer-based method achieves an 81% in precision and 44% in salience, while the order-oriented inference improves both the precision and salience, to 91% and 54%, respectively.

5.3 Salient negations in Loosely Structured KBs

While encyclopedic KBs such as Wikidata, YAGO, and DBpedia are well-canonicalized, commonsense KBs such as ConceptNet [116] and Ascent [78, 79] express information using uncanonicalized short phrases. For instance, Ascent contains the actions *lay eggs*, *deposit eggs*, and *lie their eggs*. Therefore, the methods designed for well-structured KBs will result in many incorrect inferences, e.g., $\langle \text{butterfly, capableOf, lie their eggs} \rangle$ but $\neg \langle \text{butterfly, capableOf, lay eggs} \rangle$ and $\neg \langle \text{butterfly, capableOf, deposit eggs} \rangle$. Moreover, the use of the PCA rule (only infer an absent object to be negative in the presence of at least one other object for the same predicate) to improve accuracy of inferred negations would not be sufficient, as most commonsense KBs rely on ConceptNet's well-defined, but few and generic predicates, e.g., *hasProperty*. As opposed to Wikidata's *citizenOf* predicate where only a few objects are expected, ConceptNet's *hasProperty* can have hundreds of accepted object phrases. Hence, it is not a very good idea to assume that since an entity has at least one general property present the list of objects is complete.

NegatER [107, 108] is a recent method for identifying salient negations in commonsense KBs. Given a subject s (an everyday concept in this case) and an input KB, a pre-trained **language model (LM)** is fine-tuned using the training and testing sets of KB statements. The positive statements are simply queried from the KB, while negative samples are randomly generated under the closed-world assumption, i.e., the KB is complete and the negations are generated by corrupting parts of positive statements with any other random concept or phrase. The LM is then trained to learn a decision threshold per predicate, and hence, the fine-tuned LM serves as a true/false classifier for unseen statements. To create the set of *informative* or *thematic* negations for s , s is replaced by a neighboring *entity phrase* from the KB.

Example: $\langle \text{horse, isA, pet} \rangle$ is replaced by the neighbor subject *horse rider*, resulting in the candidate negation $\neg \langle \text{elephant, isA, pet} \rangle$.

The classifier (LM) then decides on the *falseness* of such candidate. Once the set of candidates is constructed, they are finally ranked using the fine-tuned LM by descending order of negative likelihood. Even though NegatER compiles lists of thematically relevant negatives, one major limitation is that it generates many type-inconsistent statements, due to the absence of a taxonomy, e.g., $\neg \langle \text{horse rider, isA, pet} \rangle$.

Uncommonsense [12] identifies salient negations about target concepts (e.g., *gorilla*) in a KB by computing comparable concepts (e.g., *zebra*, *lion*) using external structured taxonomies (e.g., WebIsALOD [47]) and latent similarity (e.g., Wikipedia embeddings [129]). Similar to the peer-based negation inference method [8], the PCWA is used to infer candidate negations (e.g., *has no tail*, *is not territorial*). A crucial difference is the technique used for checking the accuracy of candidate negations. As previously mentioned, the PCA would not be sufficient in loosely structured KBs. Therefore, this work introduces different scrutiny steps to improve the accuracy of candidates. It performs semantic similarity checks (using Reference [101]) against the KB itself and external source checks (using pretrained LMs). Semantic similarity also contributes in grouping negative phrases with the same meaning to boost their order in the final ranked list, i.e., the relaxed sibling frequency. The finally generated top-ranked negations are extended with provenances showing

why certain negations are interesting. For example, *gorilla* has no tail, unlike other land mammals such as *lions* and *zebras*. On 200 randomly sampled concepts with their top-2 negations [12], in a crowdsourcing task where workers are asked about the accuracy and interestingness of given negative statements, UnCommonSense achieves a precision of 75% and a salience of 50%, while NegatER achieves a similar precision of 74% but lower salience of 29%.

5.4 Salient Negations in Text

Large textual corpora can be good external sources for implicit and explicit negations. Moreover, due to the incompleteness of existing KBs, text-based methods can be complementary to inference-based methods.

Mining negations from query logs [8] is an unsupervised pattern-based methodology that extracts salient negations from query logs. The intuition is that users often ask why something does not hold, as in “*Why didn’t Stephen Hawking win the Nobel Prize in Physics?*” Such a query can then be used to deduce that *Hawking* did indeed not win the *Nobel Prize*. The work defines nine negated why-questions, such as *Why didn’t <s>?*

Example: Given the pattern *Why didn’t Stephen Hawking..*, the auto-completion API of a search engine, e.g., Google’s, produces *won the Nobel Prize in Physics, accept the knighthood,*

Mining negations from text is also a possibility, and although there exist encouraging proposals [115], they have not led to significant resources so far [6]. The reason appears to be that long texts (such as newspapers, blogs, and encyclopedias) rarely mention salient negations, as an analysis on the STICS [49] corpus has shown. Sentences with a negative meaning in newspapers and blogs are mostly about things that people did not do or did not say—as in “*Brad Pitt did not threaten Angelina Jolie with cash fine,*” or “*Angela Merkel never made much of an effort to ensure that eastern Germans felt a sense of belonging.*” Encyclopedias, however, mostly contain only positive statements. The few sentences that do contain negation usually contain either double negation, temporary negatives (as in “*Hawking was not initially successful academically*”), or negations of specification (as in “*His family could not afford the school fees without the financial aid of a scholarship*”). Overall, none of these sources contain short trivia sentences with negative keywords.

Mining negations from edit logs [58] is a work that exploits the edit history of collaborative encyclopedias such as Wikipedia as a rich source of implicit negations. Editors make thousands of changes every day for various reasons, including fixing spelling mistakes, rephrasing sentences, updating information on controversial topics, and fixing factual mistakes. The work focuses on mining data from the last category. In particular, it looks at sentence edits in Wikipedia where only one entity or one number is changed.

Example: “The discovery of uranium is credited to *Marie Curie*” is updated to “The discovery of uranium is credited to *Martin Heinrich Klaproth*” where the entity *Marie Curie* is replaced with *Martin Heinrich Klaproth*. The former is then considered a common mistake, i.e., an interesting negative statement.

To decide whether an update must be labeled as *common factual mistake* or one of the other categories, a number of heuristics are applied. These include (i) checking how often sentence is being updated (to exclude controversial topics where different editors have different opinions); (ii) computing the edit distance between the entities (to exclude spelling corrections); and (iii) checking for synonyms (to exclude simple rephrasing of the same statement). It remains to be checked whether the edit removes or introduces a false statement. This is done by counting the number of supporting statements on the web.

Table 4. Comparison of Different Works on Salient Negation in KBs

Source	Work	Focus	Strengths	Limitations
KB	Arnaout et al. [8, 9, 11]	Interesting negations about encyclopedic entities using peer-based statistical inferences.	Subject recall. Salience due to peer frequency measures.	Precision due to KB incompleteness & modeling issues. Beyond simple negations (conditional negations).
Query logs	Arnaout et al. [8]	Interesting negations about encyclopedic entities using pattern-based query extraction.	Precision due to high-quality search engine query logs.	Subject recall due to APIs access limit.
Edit logs	Karagiannis et al. [58]	Common factual mistakes using mined textual change logs.	Precision due to heuristics including web hits computation.	Focus on precision over salience. Mined negations require canonicalization.
KB	Arnaout et al. [12]	Informative negations about everyday concepts using fine-tuned LMs.	Salience due to comparable concepts. Interpretable results through provenance generation. Can handle non-canonicalized KBs.	Recall depends on presence of subject in the external taxonomy.
KB and LM	Safavi et al. [107, 108]	Informative negations about everyday concepts using comparable taxonomic siblings.	Recall through corruptions using phrase embeddings.	Plausibility due to taxonomy not being considered.

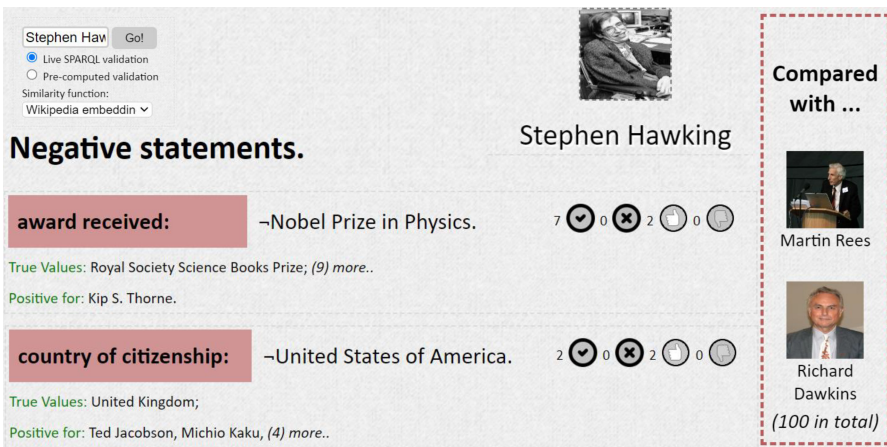


Fig. 6. User interface of the WikiNegata (<https://d5demos.mpi-inf.mpg.de/negation>) platform. It shows automatically computed salient negative statements for Stephen Hawking, such as that he did not win the Nobel Prize in Physics, unlike his colleague and friend Kip Thorne.

5.5 Summary

In this section, we defined the task of identifying salient negations about KB subjects and presented different approaches to tackle this problem. A summary of these approaches with their focus, strengths, and limitations is shown in Table 4.

Applications. Wikinegata [9, 11] is a tool for browsing more than 600M negations about 0.5M Wikidata entities. It gives insights into the peer-based method [8], where users can inspect different peers used to infer certain negations (see Figure 6). In commonsense KBs, the Uncommonsense system [12] provides a similar experience about everyday concepts.

Open challenges. For text-based methods, the main issue seems to be the subject recall. Often, negations are only expressed when they are highly exceptional and about prominent entities. For KB-based methods, the main problem is the precision-salience tradeoff. It is quite simple to get a near perfect precision when assuming the CWA, as the majority of inferred negations would be correct but nonsensical, e.g., *Stephen Hawking's capital is not Paris*. According to Reference [12], this baseline receives a 93% in precision but less than 7% in salience. As shown in this section, it becomes challenging to increase salience while maintaining a high level of precision, as plausible candidates tend to be harder to scrutinize, especially in commonsense KBs, e.g., *Is Basketball hard to learn?* Moreover, negation generators have to deal with real-world changes. This is especially true for the encyclopedic KBs where new information is added frequently. For example, prior to 2016, it was interesting that *Leonardo DiCaprio* never received an Oscar (the negation is no longer correct). A second challenge is the class hierarchy for both entity peer measures and negation generation. For example, noisy taxonomies would result in irrelevant peers. There are also modeling issues and inconsistencies that most web-scale KBs suffer from, especially the collaborative ones. For instance, to express that a person is vegan, should the editors use $\langle person, lifestyle, veganism \rangle$ or $\langle person, isA, vegan \rangle$. While one can be asserted, the other could be mistakenly negated by one of the discussed methods. Finally, the methods presented in this section are meant to compile a list of simple negative statements about commonsense and encyclopedic entities. Complex negatives, however, need further investigation and are more challenging. In scientific knowledge, for instance, two contradictory facts might be true under different contexts, e.g., *water cannot extinguish every type of fire*, such as *petrol fires*, but that does not mean that *water cannot extinguish fire*. Also, in socio-cultural knowledge, the same statement can be both true and false under different cultural factors, e.g., *drinking wine at weddings* (in *Europe* v. the *Middle East*).

LLMs for Negation Generation. Very recent studies examined the ability of **large language models (LLMs)**, such as chatGPT [82], to generate salient negative statements [7, 22]. Findings in Reference [22] are that contradictions exist in the LLM's belief, when comparing results of different tasks targeting the same piece of knowledge. For instance, LLMs generate the sentence "*Lions live in the ocean*" but answer "No" when asked "*Do lions live in the ocean?*" Lessons from Reference [7] include the importance of prompt engineering in this task. Prompts with expressions "negative statements," "negated statements," and "negation statements" return very different types of responses. Moreover, LLMs struggle with the true negativity of the statements returned, often generating statements with negative keywords but a positive meaning, e.g., "a coffee table is not only used indoors."

6 RELATIVE RECALL

So far, the yardstick for recall/completeness has been the real world: *How many of the entities (or predicates or statements) of the domain of interest have been captured?* While this generally is a meaningful target, in some cases, the notion is not well-defined or not informative. In the following, we look at alternative formulations by relaxing the absolute yardstick into a relative one: *Relative to other entities/resources/use cases, how much is the KB's recall?* We look at this problem in three variants: (i) recall within the same KB, relative to other entities, (ii) recall relative to other resources such as KBs or texts and (iii) recall relative to the extrinsic use case of question answering (see Figure 7).

6.1 Entity Recall Relative to Other Entities

We define the problem of relative entity recall as follows:

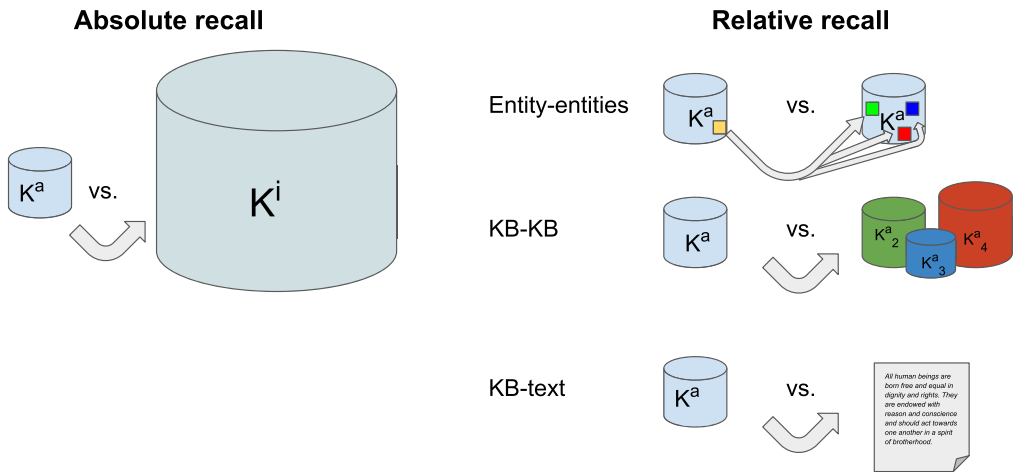


Fig. 7. Difference between absolute recall (left) that was discussed in previous sections and various notions of relative recall (right).

Definition 11. Relative Entity Recall Problem

Input: An entity e , and a KB K

Task: Determine the recall of K for statements about e , relative to related entities in K .

For example, one may ask how complete Wikidata’s information on *Albert Einstein* is, relative to similar entities. The instantiation of this problem has two major components:

- (1) How are related entities defined?
- (2) How is recall quantified and compared?

Entity relatedness is a topic with much history in data mining, and a wide range of text-based, graph-based, and embedding-based similarity measures exists (see Section 5 and, e.g., References [88, 128]). Similarly, recall can be quantified in a variety of ways, for instance, via the number of statements, predicates, inlinks, outlinks, and so on. We shall now see different approaches to both problems.

Relative recall indicators. One of the first approaches to the problem of relative entity recall is **Recoin** (**R**elative **c**ompleteness **i**ndicator) [2, 15]. It extends the entity page of Wikidata with a traffic-light-style recall indicator, indicating how comprehensive the information is compared to related entities. Its definition of relatedness is class-driven: *Paris* would be compared with other capital cities, *Radium* with other chemical elements, Albert Einstein with other Physicists (treating Wikidata’s occupations as pseudo-classes). Its quantification of recall follows a simple frequency aggregation: In each class, the top most frequent properties are computed. For instance, for capitals, the most frequent *predicates* are *country* (99%), *coordinate location* (97%), and *population* (82%). Then, the frequencies of the top-five absent properties are added up for the given entity, and the sum is compared with five global discrete thresholds to arrive at a final traffic light color.

A second approach to the problem of relative recall is provided by Wikimedia’s ORES machine learning platform [70]. The *Wikidata item quality module* specifically assigns probabilities to entities belonging to one of five quality levels (A to E), subsuming, besides recall, several other quality dimensions such as completeness, references, sitelinks, media quality. The scores are also relative in the sense that good item quality has no inherent definition, but can only be understood

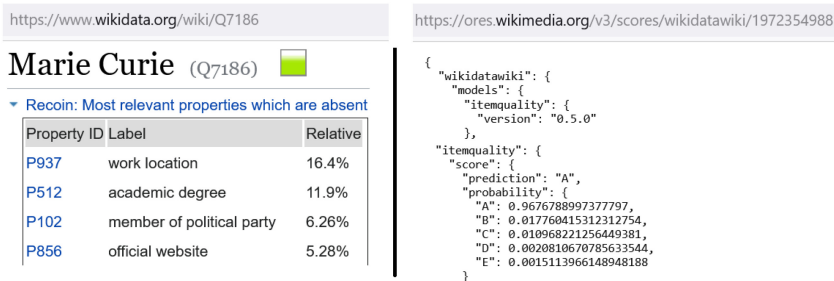


Fig. 8. Illustration of Recoin (left) and ORES (right) outputs for *Marie Curie* on Wikidata. Recoin outputs show class frequencies of absent properties, while ORES quality classes A-E are computed using a supervised regression model.

in comparison with other good/bad items. ORES apparently employs supervised machine learning based on a combination of content embeddings and latent features, yet its concrete working is only partially documented. We show example outputs of Recoin and ORES for *Marie Curie*, as of September 2023, in Figure 8.

Property ranking. Underlying both Recoin and ORES is the question of how to decide whether, for a given entity, a certain predicate is expected to be populated. This is referred to as the problem of *property ranking*. Given an entity like *Marie Curie*, is it more relevant that her doctoral advisor is recorded or the sports team she played for?

Property ranking relates to the problem of statement ranking [16, 63], though the absence of an object value makes the problem harder. The frequency-based ranking, like the one used in Recoin [15], provides a reasonable baseline, yet frequencies are only a proxy for relevance. For example, the most frequent properties typically concern generic biographic statements such as place of birth, date of birth, while interesting aspects such as scientific discoveries, awards, or political affiliations are expressed less frequently.

Property ranking for relative recall assessment has thus been advanced in several ways: In Reference [95], text-based predictive models are proposed, which are trained on Wikipedia descriptions of entities and describe how likely a textual description is accompanied by a certain structured property in Wikidata. In an evaluation of pairwise property relevance predictions, this approach achieves up to 67% agreement with human annotators. Gleim et al. introduce SchemaTree [44], a trie-based method for capturing property frequencies in the existing data, taking into account not just individual frequencies, but also frequencies of combinations and fallbacks to base cases for rare combinations. Issa et al. rely on association rule mining [54]. Luggen et al. [66] propose a method based on multimodal Wikipedia embeddings, taking into account multilingual article descriptions as well as pictures. They train a multi-label neural classifier on these embeddings, for the task of predicting currently present properties in Wikidata, and find that this significantly outperforms previous approaches.

Analysis. The discussed approaches mostly treat entities as sets of properties—distinguishing just whether a property is present or not, but not taking into account how many values are present. For example, even though one award for *Einstein* is listed in a KB, the KB could be missing many others. Disregarding this aspect is pragmatically motivated by the partial completeness assumption (see Section 2): The presence of one value per property likely implies presence of all values. At the same time, extensions towards explicit regard for multi-valued properties would be desirable. A second limitation is inherent to all relative measures: Relative measures may appear good even for bad items if the comparison is even worse and vice versa. There is no firm link to the notion

of recall from the previous sections, as, even if an entity is well-covered compared with related entities, no strong deductions about its recall, relative to the real world, are possible.

6.2 KB Recall Relative to other Resources

There are several other categories of resources that can be used for KB recall assessment. A relatively straightforward comparison are other KBs, while texts and elicited human associations provide different foci, on recall w.r.t. information typically conveyed in texts, and w.r.t. information that humans spontaneously associate with concepts.

KBs to KBs. One way to compare KB recall is by comparing their size. Virtually every KB project compares itself with other resources in terms of size, typically counting the number of entities, classes, and statements (for one example, see Table 1 in the Knowledge Vault paper [32]). A more fine-grained analysis was performed in Reference [103]. For 25 important classes, considerable variance was found, even on KBs derived from the same source, such as DBpedia and YAGO. Variances were in part explained by different modeling approaches (e.g., Wikidata’s low number of politicians explained by modeling a class via properties). However, size comparisons do not capture whether or to which degree content from one KB is contained in the other—it may well be that resources have different foci and that the larger ones still have limited recall w.r.t. smaller ones.

To avoid such issues, we can look at the fraction of entities from one contained in the other. The work in Reference [103] analyzed that as well, matching entities via simple String distance functions. Merging KBs generally provided potential for increasing recall even for the already bigger resources, because of differences in focus.

KB vs. text. Texts are a prime mode of knowledge storage and sharing, and it is natural to ask how well KBs recall information from texts. Notably, KBs were born in part out of the shortcomings of texts in terms of structuring information, so it is very interesting to investigate to which degree the higher structure of KBs comes at a loss of recall. Moreover, a principled methodology to compare KB recall w.r.t. texts enables a comparison on a much wider set of domains and subdomains than the structured comparison in the KB-to-KB setting, as texts are available in much bigger abundance than structured resources. Various approaches to estimating KB recall w.r.t. texts exist.

In the Aristo TupleKB project [73], recall of an automatically constructed science KB is compared with information contained in scientific texts. For this recall assessment, Mishra et al. assemble a corpus of 1.2M sentences from elementary science textbooks, Wikipedia, and dictionaries, from which they automatically extract relational statements. They can then quantify the “science recall” of a KB by measuring which fraction of these statements is contained in a KB. This is done for five KBs (WebChild, NELL, ConceptNet, ReVerb-15M, TupleKB). Since predicate names in the KBs vary, this analysis is restricted to 20 general predicates, which are manually mapped to a wider set of surface names. Also, for subject and object match, only headwords are considered. Under these relaxed conditions, the science recall of these KBs is found to be between 0.1% and 23.2%.

A related analysis is provided in the context of the OPIEC project [40], a corpus of open statements extracted from running an open information extraction system on the entire English Wikipedia. In Reference [39], the authors analyze the relation of OPIEC with DBpedia, in particular, to which degree OPIEC statements are expressible in DBpedia and to which degree they are actually expressed. Since open information extraction provides a plausible approximation of relational statements contained in text, this evaluation can be seen as a relative recall evaluation of DBpedia w.r.t. Wikipedia. They provide several important insights: (i) 29% of open statements can be fully expressed in DBpedia, 29% partially (e.g., with a more specific or more generic predicate),

42% of open statements are not KB-expressible.¹⁴ (ii) Adding more complex constructs, such as conjunctions or even existential quantification, into the KB increases its recall potential. (iii) When it comes to measuring the actual recall, they find that 18% of open statements are fully present in DBpedia, 23% partially present, 59% not at all.

In the temporal dimension, KB vs. text recall has been studied using the distant supervision assumption: Assuming that hyperlinked entities on Wikipedia pages are relevant to the page subject, how does their recall in Wikidata change over time? Wikidata’s revision history allows to study this longitudinal, allowing to observe a generally steady increase in recall from 2012 to 2020 [96].

KB vs. human associations. Besides other KBs and texts, an interesting resource for relative recall assessments are humans directly. How well do KBs recall statements that humans spontaneously associate with concepts? This framing of recall has been prominent in recent commonsense **knowledge base construction (KBC)** projects [77, 78, 105]. While the former project directly queried human crowdworkers (“What comes to your mind when you think of lions?”), the latter two projects relied on the CSLB property norm corpus [29], a large dataset of concept associations collected in the context of a psychology project. Evaluations following this scheme typically use embedding-based heuristic matching techniques to judge whether a test KB contains a reference statement subject to possible minor wording differences. Recall is typically found to be on the order of 5%–13%, showing that there is still a substantial gap in how well current commonsense KB projects cover human knowledge.

Analysis. We have discussed three ways of estimating KB recall relative to other resources. Common among them is the challenge of how to compare pieces of knowledge: Statements in the reference resource may be differently worded or ambiguous, making finding a statement of equivalent semantics nontrivial. Moreover, semantic equivalence is necessarily an imprecise concept, requiring somewhat arbitrary decisions about the maximally accepted dissimilarity, as well as difficult technical decisions on how to actually measure semantic relatedness. A common technical solution currently is to represent the two statements to be compared in latent embedding space (e.g., Reference [77] uses S-BERT [101]), then use similarity metrics like cosine distance to decide whether two candidates are considered equivalent. Note that this problem even occurs in the most structured setting, KB-vs.-KB recall, since even if subjects and objects are disambiguated, predicate names are typically textual strings, such as *worksAt*, *employer*, *affiliatedWith*.

Each of the mentioned comparisons serves a different purpose: Evaluating KB recall relative to other KBs should definitely be attempted whenever similar-domain KBs are available, since data integration from structured sources is a comparatively simple task, with considerable potential for improving recall. Evaluating KB recall relative to text can serve two purposes: Towards improving KB recall, for text-extracted KBs, recall evaluation relative to text can serve as a guidance on how much information is lost in the text extraction process. This can serve as a guidance on where to improve the process. However, such an evaluation can also help to understand a KB’s potential and limitation for downstream use cases (e.g., Wikidata is generally suited for computing statistics about age/gender/profession statistics about US senators, but bad for judging the quality of their governance). This is especially relevant when KBs are one of several options to power a downstream use case, as is, for example, often the case for question answering (see also Section 6.3). KB recall relative to human associations is the most intrinsic dimension.

¹⁴This aspect can also be considered another dimension of recall, called *schema recall*. This aspect is as also studied in Reference [76], where the authors find that popular KB schemata contain between 13–126 predicates of a sample of 163 predicates found in a few Wikipedia articles.

6.3 KB Recall Relative to Question Answering Needs

Another relative way to evaluate KB recall is by considering a use case and quantifying to which degree KBs satisfy its data needs. Natural-language *question answering* is arguably a supreme task in knowledge management and NLP and as such especially suited to illustrate this point.

Static analyses. Several papers investigate to which extent KBs allow answering questions from common QA datasets. Most prominently, this happens in scenarios where text-based QA systems and KB-based QA systems are compared. Two especially illustrative comparisons can be found in References [81] and [89]. The first work analyzes the performance of a state-of-the-art QA system on four popular benchmarks of general-world questions (NaturalQuestions, WebQuestions, TriviaQA, CuratedTREC). It finds that KB-based systems can correctly answer between 26% and 43% of these queries. It also provides a comparison of KB recall to text recall (see Section 6.2) by comparing the previous numbers with the performance of text-based QA systems (finding 45%–62% recall). In other words, texts generally provide higher recall than KBs, although this needs to be weighed against other advantages of structured resources. The work of Reference [89] analyzes complex queries in six QA benchmark datasets (LCQuad 1.0 and 2.0, WikiAnswers, Google Trends, QuALD, ComQA), finding that a state-of-the-art system can answer 10%–19% of these queries by using only KBs, and 18%–36% by only using text. Further variants of this kind of analysis exist, for example, for conversational question answering [23].

The common insight from these studies is that KBs provide encouraging recall for utilizing them in QA systems, but are far from saturating the query sets, thus often motivating hybrid QA systems that combine KBs and text. At the same time, the reported scores typically conflate the recall of the KB and the ability of the system to pull out the correct answers from the KB. Since query answering is a heuristic, imperfect process, the intrinsic KB recall for these QA datasets is likely higher.

Predicting recall requirements for the QA use-case. An interesting twist to recall analysis is provided by Hopkinson et al. [52]: Instead of assuming a fixed KB and measuring recall, they define a desired query recall (95% of queries should be answered by the KB) and ask which content needs to go into the KB to achieve this. This is not an obvious question, because information needs may vary highly depending on the type of entity, and questions do not uniformly target entities and properties. The work originates in an industrial lab (Amazon Alexa), where such a business requirement is plausible. The QA service provider here has the potential to design automated extraction efforts accordingly or to task paid KB curators to complete specific areas of the KB. Moreover, commercial service providers have access to user query logs, which underlie this technique. On a technical basis, the work represents entities via their class membership, extracts usage frequencies of properties per entity from the query log, and predicts predicate usage patterns on new entities using either a regression or a neural network model. Results indicate that this method can predict required properties with good accuracy. Furthermore, demand-weighted requirements can be lifted to the level of the whole KB, based on usage data about which entities are queried how often. For a non-representative set of entities in the Alexa KB, the authors find that 58% of the predicates needed to arrive at the 95% query-answerability goal are currently in the KB.

Longitudinal development. Reference [96] analyzes how KB utility has changed over time. The authors select questions from three search engine logs (AOL, Google, Bing queries), then use human annotation to find out the earliest time at which a KB (Wikidata and DBpedia) could answer these questions. For example, the question “*Where is Italian Job filmed?*” can be answered by Wikidata since October 15, 2015, when the property *filming location* on the entity *Italian*

Table 5. Summary of Relative Recall Assessment

Focus of relative recall	Aspect	Works	Strength	Limitations
Entity vs. other entities	Relative entity recall	[15, 70]	Enables ranking/prioritization without knowledge about reality	Struggles if many properties are optional/if only parts of values are present
	Determining most relevant predicates per entity	[44, 66, 95]	Produces interpretable suggestions on where to complete KB	Struggles with optional predicates
KB to other resources	KB to KB	[32, 103]	Can give quick suggestions on when to integrate data	Similar-topic KBs often not available
	KB to text	[39, 40, 73, 96]	Can help identifying issues in text extraction or help in choices between KB or text-based downstream applications	Threshold of when textual statement is covered in KB not obvious
	KB to human associations	[77, 78, 105]	Highest aspiration of all evaluations	Practical implications not clear
KB to QA use cases	Counting #queries answered by KB-QA system	[23, 81, 89]	Gives tangible insights into how well KB feeds a use case	Difficult to disentangle KB recall and QA system performance
	Predictive QA recall	[52]	Allows to predict content needed in KB for meeting a use case requirement	Requires substantial query logs
	Longitudinal development	[96]	Shows that KBs have steadily improved for QA	Relies on heuristic matches

Job was added. Plotting the number of answerable queries over time turns out to show a steady increase for the time period from 2003 to 2020, with only minor slowing in recent years.

6.4 Summary

Recall estimation in absolute terms is generally a difficult task. This section has provided a pragmatic alternative, showing how to measure KB entity and statement recall relative to other KBs, other resources, and QA use cases.

We summarize the insights from this section in Table 5. Each of these approaches comes with advantages and disadvantages, mostly stemming from challenges in how to measure relative recall, and the potential of systematic omissions in the reference.

Relative entity recall is arguably the easiest to analyze, and the quantification is comparatively easy, as matching predicates within a KB is simple. Thought is still required in the definition of the reference entities, e.g., comparing Einstein with other Physicists may give very different results than comparing him with other violin players. Furthermore, relative entity recall is prone to systematic gaps in the KB, e.g., if the predicate *known for* is entirely absent from the KB, then there is no way to suggest it for Einstein either.

Relative recall to other resources provides a more external view of a KB, and especially when choosing text as reference, provides potential for many interesting analyses (e.g., recall of Einstein's KB entry can be computed w.r.t. Wikipedia, w.r.t. Simple Wikipedia, w.r.t. a biographical book). At the same time, quantifying recall w.r.t. external resources is more challenging, as it requires dealing with schema matching (KB-KB setting) or imprecise and ambiguous predicate and entity surface forms (KB-text setting).

Relative recall w.r.t. use cases provides clear metrics of how a KB is faring downstream, and this is advantageous if a KB is built with a primary use case in mind. This strength is, however, often also a challenge, since, often, KB construction is a longitudinal, cross-functional endeavor [127], where use cases are moving targets.

7 DISCUSSION AND CONCLUSION

We conclude this survey with a discussion of the temporal dimension of recall (Section 7.1), the impact of large language models (Section 7.2), recommendations on how to perform KB recall

assessment (Section 7.3), a set of take-home lessons (Section 7.4), and list of challenges for future research (Section 7.5).

7.1 The Temporal Dimension of Completeness and Recall

Reality continuously changes: People who were once presidents lose that role, people marry or divorce, and people who once did not have a death date may obtain one. Knowledge bases follow this course and are usually updated. In this section, we discuss the impact that these changes have on the tasks of completeness and recall estimation.

Many KBs grow steadily. Wikidata, for instance, contained 20M statements in 2015, but 1.2B statements in 2023, hinting at a substantial increase of its recall. Reality, however, evolves as well, which may mean that areas formerly complete might become incomplete later on. In this section, we discuss formalisms for temporal annotations, methods for extraction and extrapolation, and observational studies concerning recall trends.

Formalisms for Temporal Annotations. Completeness and cardinality statements can be extended with information on their temporal validity. Darari et al. [27], for instance, introduced the notion of *time-stamped completeness statements*. These add a “latest validity” date to a completeness statement (which is typically, but not necessarily, the date of their creation). Examples are statements such as “*Nobel Prize winners are complete until 2023*” or “*XYZ’s publications are complete until 2018*.” Similarly, Arnaout et al. [10] extended negative statements inference with a notion of temporal prefix, allowing to conclude, for instance, that unlike her seven predecessors, the German chancellor Angela Merkel was not male.

In many KBs, temporal annotations are also used for positive statements. A typical statement, in Wikidata, for instance, qualifies that Albert Einstein’s German citizenship ended in 1933 or that he received the Nobel Prize in Physics in 1921.¹⁵ Annotating completeness and recall statements with temporal qualifications appears therefore quite natural.

Extracting and extrapolating time information. For completeness and recall information that is text-extracted, a reasonable baseline is to consider the extraction time as the latest validity time. Finer-grained annotations are possible, for instance, by considering document creation time metadata or temporal expressions in the actual text for estimating the latest (certain) validity time [117].

Given time-annotated completeness or recall metadata, an orthogonal question is how to interpret it after its latest validity time. For instance, if publications by Albert Einstein were complete in 2018, then it is reasonable to assume that the same is the case in 2023. Yet, no such conclusion should be drawn for books *about* Albert Einstein. Drawing this distinction means entering the realm of predicting the temporal stability of knowledge, a problem that comes with a dependence on domain knowledge and which appears under-explored for both structured knowledge [31, 94] and text [4].

Recall trends. KBs such as Wikidata, YAGO, and DBpedia are under active development, and one may wonder how their recall evolves. In Section 6.3, we discussed work that analyzed recall relative to QA needs, and Wikipedia content, for DBpedia and Wikidata [96]. Remarkably, relative to fixed information needs (QA logs from the 2000s, a snapshot of Wikipedia pages), recall had constantly increased from 2003–2020, with only a minor slowdown in recent years. However, reality and information needs also evolve over time, and hence it remains an open question of whether KBs are faster in capturing reality or reality develops faster than KBs can represent it.

¹⁵<https://www.wikidata.org/wiki/Q937>

7.2 The Impact of Large Language Models

Recently, **large language models (LLMs)** such as BERT [30], (chat-)GPT [92], and LLaMA [122] have significantly impacted natural language processing. This impact has extended to knowledge-intensive tasks and specifically also to knowledge bases [83, 120]. Although completeness and recall research has yet to capitalize on these advances, there are several ways by which LLMs are likely to impact this area.

Indirect impact: KBs with higher recall. LLMs support many steps in the KB construction pipeline, thereby enabling the construction of bigger KBs that consequently have higher recall [3, 124]. In particular, LLMs can be used both for direct knowledge prediction or in conjunction with retrieved text, where they improve over existing textual relation extraction methods [68].

Direct impact: Easier linking of text and structured modalities. Linking textual statements with structured statements is a problem that affects several of the discussed methodologies, most notably peer-based inferences (see Section 5, where existing statements need to be matched with statements on peers and with textual evidence), and the KB-to-text relative recall assessment (Section 6.2). Due to their significant capabilities in latently representing and matching assertions in different formulations and representations, we may expect advances on these parts soon.

Direct impact: Conversational maxims. LLMs are especially strong in generalizable linguistic tasks that do not require instance-specific knowledge. In the context of recall assessment, one such task concerns the estimation of whether a textual extraction context implies completeness or not (see Section 3.2). For example, chatGPT can give convincing answers to the following examples that come with very subtle linguistic nuances:

User: Text entailment: Please estimate how likely the first sentence entails the second.
 Premise: John brought his daughter Mary to school.
 Hypothesis: John has only one child.

chatGPT: [...] high likelihood.

User: Text entailment: Please estimate how likely the first sentence entails the second.
 Premise: John also brought his daughter Mary to school.
 Hypothesis: John has only one child.

chatGPT: [...] does not entail.

Direct impact: Generating metadata. Mirroring approaches that directly prompt LMs for statements [3, 124], LLMs can, in principle, also be prompted for completeness or recall metadata. An example is shown next.

User: According to Wikidata, Barack Obama has two children, Malia and Sasha. Are these truly all?

chatGPT: Yes, according to publicly available information, Barack Obama has two daughters [...].

User: According to Wikidata, Albert Einstein has published 75 scholarly articles. Are these truly all?

chatGPT: No, the number of scholarly articles published by Albert Einstein is estimated to be around 300. [...]

In the example above, chatGPT gives correct answers to both questions, confirming completeness for Barack Obama's children and rejecting completeness for Einstein's publications, while also

giving recall information (75/300). However, metadata obtained in this way itself comes with huge credibility issues. LLMs are known to confabulate, especially for long-tail topics, so without a proper sourcing of answers, such derivations are likely difficult to utilize downstream. Sourcing LLM generations in texts is difficult for principled reasons and an open research challenge [131]. One exception to this general issue may be negative commonsense knowledge (cf. Section 5.3). Here, correctness depends anyway much on context, and requirements in correctness are balanced by requirements for saliency. In this case, LLMs can reasonably generate interesting negation candidates [7].

7.3 Recommendations for KB Recall Assessment

We have aimed for a balanced coverage of approaches so far, which may leave practically minded readers wondering how they could best approach a specific KB recall assessment problem. In the following, we give more concrete suggestions on how we think a sensible order of approaches to specific problems could be. We distinguish three settings: (i) open-ended ab-initio KB construction, where recall-awareness can be intertwined with construction efforts; (ii) use-case driven construction, where efforts can be directly matched with use case metrics; and (iii) KB curation, where an existing KB shall be evaluated.

Setting 1: Open-ended ab-initio KB construction. For open-ended ab-initio KB construction, i.e., the novel construction of a KB intended for broad use, our suggestion is to intertwine the data acquisition process with metadata acquisition. Concretely, if the KB content is text-extracted, we suggest to use the text recall estimation techniques from Section 3.2, to annotate extractions with confidence values concerning their completeness. If the data is created by human authors, we suggest to augment the data authoring tools with fields for metadata collection, e.g., checkboxes that allow authors to note when they finished recording a topic, similar as in Cool-WD [90].

Setting 2: Use-case-driven ab-initio KB construction. For KB construction driven by a specific use case, our suggestion is to organize the recall assessment via metrics derived from the use, as discussed in Section 6.3. Initially, one should derive a profile of queries to be answered by the KB, e.g., by sampling from the use case. Where the sample's breadth is limited, interpolation should be used to derive a broader profile [52]. Efforts towards KB population can then be evaluated against this query profile, i.e., for each specific population technique or domain, one could compare cost and benefit and prioritize accordingly.

Setting 3: KB curation. In settings where existing KBs shall be evaluated, we suggest to first check for the existence of KB-internal cardinality information, as discussed in Section 4.1. Next, if high-quality texts like Wikipedia are available, then we suggest to exploit textual cardinality assertions, preferably with simple template-based extraction, as in Section 4.2. Relative recall, in particular, by comparing entities inside the KB (see Section 6.1), will also help in spotting gaps. Statistical properties such as the ones we have discussed in Section 3 should only be used once all other options have been exhausted, because they are least reliable.

7.4 Take-home Lessons

The key takeaways from this survey are:

- (1) **KBs are incomplete:** Despite the long history of the fields of KB construction, Semantic Web, and Information Extraction, the construction of a general-world knowledge base is an inherently fuzzy and evolving task. Therefore, such KBs will always be incomplete, and one has to be able to deal with this incompleteness, rather than hoping that it will disappear (Section 2).

- (2) **KBs do hardly contain negative information (but should):** Negative information is very useful for downstream tasks, but regrettably underrepresented in current KBs. Selective materialization of interesting negations can significantly enhance the utility of KBs (Sections 2 and 5).
- (3) **Predictive techniques work for a surprising set of paradigms:** Besides classical supervised prediction, there are statistical properties such as number distributions, sample overlap, and density invariants that enable recall prediction even without typical training (Section 3).
- (4) **Count information is a prime way to gain insights into recall:** Count information provides the most direct way to recall assessment, and it can be found both in existing KBs and in text (Section 4).
- (5) **Salient negations can be heuristically materialized:** Although negative knowledge is quasi-infinite, heuristics for materializing relevant parts can significantly complement positive-only KBs (Section 5).
- (6) **Relative recall is a tangible alternative to absolute notions:** Comparing KB entities with other KB entities, external resources, or use-case requirements provides a valuable second view on KB recall (Section 6).

7.5 Challenges and Opportunities

In this final section, we sketch some of the open challenges that remain to be addressed towards fully understanding KB recall, pointing out opportunities for original and potentially impactful research.

- (1) **Developing high-accuracy recall estimators.** Most of the estimators presented in this survey are proofs-of-concept, tested only in limited domains or under very specific assumptions. Building practically usable high-accuracy estimators, possibly by combining several complementary estimation techniques, remains a major open challenge.
- (2) **Exploiting recall estimates for value-driven KB completion.** Despite their obvious connection, research on recall estimation and KB completion has so far evolved largely independently. Quantifying the value of knowledge (as in Section 6.3) and defining prioritization strategies for recall improvement that maximize the value of the available knowledge [52, 94] are great opportunities for practically impactful research.
- (3) **Estimating the recall of pre-trained language models.** Knowledge extraction from pre-trained language models has recently received much attention [87], yet it remains unclear to which degree this approach can yield knowledge for multi-valued, optional, and long-tail predicates [56, 112]. Systematically measuring the recall of language models and comparing it with structured KBs is an open challenge.

Knowledge bases have received substantial attention in recent years, and while precision is usually in the focus of construction, understanding their recall remains a major challenge. In this survey, we have systematized major avenues towards KB recall assessment and outlined practical approaches and open challenges. We hope this survey will inspire readers to reflect on KB quality from a new angle and lead to more KB projects that systematically record and reflect on their recall.

REFERENCES

- [1] Maribel Acosta, Elena Simperl, Fabian Flöck, and Maria-Esther Vidal. 2017. Enhancing answer completeness of SPARQL queries via crowdsourcing. *J. Web Semant.* 45 (2017), 41–62.
- [2] Albin Ahmeti, Simon Razniewski, and Axel Polleres. 2017. Assessing the completeness of entities in knowledge bases. In *The Semantic Web: ESWC 2017 Satellite Events (ESWC'17)*, 7–11.

- [3] Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. Prompting as probing: Using language models for knowledge base construction. In *Semantic Web Challenge on Knowledge Base Construction from Pre-Trained Language Models (LM-KBC)*.
- [4] Axel Almqvist and Adam Jatowt. 2019. Towards content expiry date determination: Predicting validity periods of sentences. In *Advances in Information Retrieval: 41st European Conference on IR Research (ECIR'19)*, Springer International Publishing, 86–101. https://doi.org/10.1007/978-3-030-15712-8_6
- [5] Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 1388–1398.
- [6] Hiba Arnaout. 2023. *Enriching Open-world Knowledge Graphs with Expressive Negative Statements*. Ph.D. Dissertation.
- [7] Hiba Arnaout and Simon Razniewski. 2023. Can large language models generate salient negative statements? In *Knowledge Base Construction from Pre-trained Language Models Workshop at ISWC 2023 (KBC-LM)*.
- [8] Hiba Arnaout, Simon Razniewski, and Gerhard Weikum. 2020. Enriching knowledge bases with interesting negative statements. In *Automated Knowledge Base Construction (AKBC'20)*.
- [9] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2021. Negative knowledge for open-world Wikidata. In *Companion Proceedings of the Web Conference 2021 (WWW'21)*. Association for Computing Machinery, 544–551. <https://doi.org/10.1145/3442442.3452339>
- [10] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2021. Negative statements considered useful. *J. Web Semant.* 71, C (2021). DOI: <https://doi.org/10.1016/j.websem.2021.100661>
- [11] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2021. Wikinegata: A knowledge base with interesting negative statements. *Proc.VLDB Endow.* 14, 12 (2021), 2807–2810. <https://doi.org/10.14778/3476311.3476350>
- [12] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2022. UnCommonSense: Informative negative knowledge about everyday concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM'22)*, Association for Computing Machinery, 37–46. <https://doi.org/10.1145/3511808.3557484>
- [13] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web (ISWC)*, Springer, Berlin Heidelberg, 722–735.
- [14] Franz Baader, Diego Calvanese, Deborah McGuinness, Peter Patel-Schneider, Daniele Nardi, et al. 2003. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press.
- [15] Vevoke Balaraman, Simon Razniewski, and Werner Nutt. 2018. Recoin: Relative completeness in Wikidata. In *Companion Proceedings of the The Web Conference 2018 (WWW'18)*. International World Wide Web Conferences Steering Committee, 1787–1792. <https://doi.org/10.1145/3184558.3191641>
- [16] Hannah Bast, Björn Buchhold, and Elmar Haussmann. 2015. Relevance scores for triples from type-like relations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. Association for Computing Machinery, 243–252. <https://doi.org/10.1145/2766462.2767734>
- [17] Mohamed Ben Ellefi, Zohra Bellahsene, John G. Breslin, Elena Demidova, Stefan Dietze, Julian Szymański, and Konstantin Todorov. 2018. RDF dataset profiling: a survey of features, methods, vocabularies and applications. *Semantic Web* 9, 5 (2018), 677–705.
- [18] Frank Benford. 1938. The law of anomalous numbers. *Proc. Am. Philos. Societ.* (1938).
- [19] Michael K. Bergman. 2001. White paper: The deep web: Surfacing hidden value. *J. Electron. Publish.* (2001).
- [20] Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2 (AAAI'07)*. <https://dl.acm.org/doi/abs/10.5555/1619797.1619981>
- [21] Andrew Carlson, Justin Betteridge, Bryan Kiesel, Burr Settles, Estevam Hruschka, and Tom Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*. AAAI Press, 1306–1313.
- [22] Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023. Say what you mean! Large language models speak too positively about negative commonsense knowledge. In *ACL*.
- [23] Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Conversational question answering on heterogeneous sources. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22)*. Association for Computing Machinery, 144–154. <https://doi.org/10.1145/3477495.3531815>
- [24] Cuong Xuan Chu, Simon Razniewski, and Gerhard Weikum. 2021. KnowFi: Knowledge extraction from long fictional texts. In *Automated Knowledge Base Construction (AKBC)*. 1–19.
- [25] Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. Association for Computing Machinery, 355–366. <https://doi.org/10.1145/2488388.2488420>

- [26] Fariz Darari, Werner Nutt, Giuseppe Pirro, and Simon Razniewski. 2013. Completeness statements about RDF data sources and their use for query answering. In *The Semantic Web. ISWC 2013: 12th International Semantic Web Conference (ISWC'13)*, Springer, 66–83. https://doi.org/10.1007/978-3-642-41335-3_5
- [27] Fariz Darari, Werner Nutt, Giuseppe Pirro, and Simon Razniewski. 2018. Completeness management for RDF data sources. *Trans. Web* 12, 3 (2018). DOI: <https://doi.org/10.1145/3196248>
- [28] Marc Denecker, Álvaro Cortés-Calabuig, Maurice Bruynooghes, and Ofer Arieli. 2008. Towards a logical reconstruction of a theory for locally closed databases. *Trans. Data Syst.* 35, 3 (2008). DOI: <https://doi.org/10.1145/1806907.1806914>
- [29] Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (CSLB) concept property norms. *Behav. Res. Meth.* 46 (2014), 1119–1127. DOI: <https://doi.org/10.3758/s13428-013-0420-4>
- [30] Jacob Devlin, Ming-Wei Chang, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [31] Ioannis Dikeoulas, Jannik Strötgen, and Simon Razniewski. 2019. Epitaph or breaking news? Analyzing and predicting the stability of knowledge base properties. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW'19)*. Association for Computing Machinery, 1155–1158. <https://doi.org/10.1145/3308560.3314998>
- [32] Xin Luna Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmman, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. Association for Computing Machinery, 601–610. <https://doi.org/10.1145/2623330.2623623>
- [33] Xin Luna Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wiko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. 2014. From data fusion to knowledge fusion. *Proc. VLDB Endow.* 7, 10 (2014). DOI: <https://doi.org/10.14778/2732951.2732962>
- [34] Patrick Ernst, Amy Siu, and Gerhard Weikum. 2015. KnowLife: A versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinf.* 16 (2015), 1–13.
- [35] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. 2018. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semant. Web J.* 9, 1 (2018), 77–129. <https://doi.org/10.3233/SW-170275>
- [36] Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M. Suchanek. 2017. Predicting completeness in knowledge bases. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM'17)*. Association for Computing Machinery, 375–383. <https://doi.org/10.1145/3018661.3018739>
- [37] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. 2015. Fast rule mining in ontological knowledge bases with AMIE+. *VLDB J.* 24, 6 (2015), 707–730. <https://doi.org/10.1007/s00778-015-0394-1>
- [38] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. 2013. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. Association for Computing Machinery, 413–422. <https://doi.org/10.1145/2488388.2488425>
- [39] Kiril Gashteovski, Rainer Gemulla, Bhushan Kotnis, Sven Hertling, and Christian Meilicke. 2020. On aligning OpenIE extractions with knowledge bases: A case study. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics, 143–154. <https://doi.org/10.18653/v1/2020.eval4nlp-1.14>
- [40] Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, and Rainer Gemulla. 2018. OPIEC: An open information extraction corpus. In *Automated Knowledge Base Construction (AKBC)*.
- [41] Shrestha Ghosh, Simon Razniewski, and Gerhard Weikum. 2020. CounQER: A system for discovering and linking count information in knowledge bases. In *The Semantic Web: ESWC 2020 Satellite Events: ESWC 2020 Satellite Events (ESWC'20)*. Springer-Verlag, 84–90. https://doi.org/10.1007/978-3-030-62327-2_15
- [42] Shrestha Ghosh, Simon Razniewski, and Gerhard Weikum. 2020. Uncovering hidden semantics of set information in knowledge bases. *J. Web Semant.* 64 (2020). DOI: <https://doi.org/10.1016/j.websem.2020.100588>
- [43] Arnaud Giacometti, Béatrice Markhoff, and Arnaud Soulet. 2019. Mining significant maximum cardinalities in knowledge bases. In *The Semantic Web – ISWC 2019: 18th International Semantic Web Conference*. Springer-Verlag, 182–199. https://doi.org/10.1007/978-3-030-30793-6_11
- [44] Lars C. Gleim, Rafael Schimassek, Dominik Hüser, Maximilian Peters, Christoph Krämer, Michael Cochez, and Stefan Decker. 2020. SchemaTree: Maximum-likelihood property recommendation for Wikidata. In *European Semantic Web Conference (ESWC'20)*. Springer, 179–195.
- [45] Herbert P. Grice. 1975. Logic and conversation. In *Speech Acts*. Brill.
- [46] Chadi Helwe, Chloé Clavel, and Fabian Suchanek. 2021. Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *International Conference on Automated Knowledge Base Construction (AKBC)*.

- [47] Sven Hertling and Heiko Paulheim. 2017. WebIsALOD: Providing hypernymy relations extracted from the web as linked open data. In *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference (ISWC)*. Springer-Verlag, 111–119. https://doi.org/10.1007/978-3-319-68204-4_11
- [48] Wassily Hoeffding. 1994. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, Springer, 409–426. https://doi.org/10.1007/978-1-4612-0865-5_26
- [49] Johannes Hoffart, Dragan Milchevski, and Gerhard Weikum. 2014. STICS: Searching with strings, things, and cats. In *Conference on Research and Development in Information Retrieval*.
- [50] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Comput. Surv.* 54, 4 (2021). <https://doi.org/10.1145/3447772>
- [51] Bernhard Hollunder and Franz Baader. 1991. Qualifying number restrictions in concept languages. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR’91)*. Morgan Kaufmann Publishers Inc., 335–346.
- [52] Andrew Hopkinson, Amit Gurdasani, Dave Palfrey, and Arpit Mittal. 2018. Demand-weighted completeness prediction for a knowledge base. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers) (NAACL’18)*. Association for Computational Linguistics, 200–207. <https://doi.org/10.18653/v1/N18-3025>
- [53] Subhi Issa, Onaopepo Adekunle, Fayçal Hamdi, Samira Si-Said Cherfi, Michel Dumontier, and Amrapali Zaveri. 2021. Knowledge graph completeness: A systematic literature review. *IEEE Access* 9 (2021), 31322–31339. <https://doi.org/10.1109/ACCESS.2021.3056622>
- [54] Subhi Issa, Pierre-Henri Paris, and Fayçal Hamdi. 2017. Assessing the completeness evolution of DBpedia: A case study. In *Advances in Conceptual Modeling: ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ*. Springer, 238–247.
- [55] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 2 (2022), 494–514. <https://doi.org/10.1109/TNNLS.2021.3070843>
- [56] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Trans. Assoc. Computat. Ling.* 9 (2021), 962–977. https://doi.org/10.1162/tacl_a_00407
- [57] Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. GenIE: Generative information extraction. In *NAACL*.
- [58] Georgios Karagiannis, Immanuel Trummer, Saehan Jo, Shubham Khandelwal, Xuezhi Wang, and Cong Yu. 2019. Mining an “antiknowledge base” from Wikipedia updates with applications to fact checking and beyond. *Proc. VLDB Endow.* 13, 4 (2019), 561–573. <https://doi.org/10.14778/3372716.3372727>
- [59] Nicholas Klein, Filip Ilievski, Hayden Freedman, and Pedro Szekely. 2022. Identifying surprising facts in Wikidata. In *Wikidata Workshop*.
- [60] Graham Klyne, Jeremy J. Carroll, and Brian McBride. 2004. Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation.
- [61] Jonathan Lajus and Fabian M. Suchanek. 2018. Are all people married? Determining obligatory attributes in knowledge bases. In *Proceedings of the 2018 World Wide Web Conference (WWW’18)*. International World Wide Web Conferences Steering Committee, 1115–1124. <https://doi.org/10.1145/3178876.3186010>
- [62] Willis Lang, Rimma V. Nehme, Eric Robinson, and Jeffrey F. Naughton. 2014. *Partial Results in Database Systems (SIGMOD’14)*. Association for Computing Machinery, New York, NY. <https://doi.org/10.1145/2588555.2612176>
- [63] Philipp Langer, Patrick Schulze, Stefan George, Matthias Kohnen, Tobias Metzke, Ziawasch Abedjan, and Gjergji Kasneci. 2014. Assigning global relevance scores to DBpedia facts. In *2014 IEEE 30th International Conference on Data Engineering Workshops*, IEEE, 248–253.
- [64] Alon Y. Levy. 1996. Obtaining complete answers from incomplete databases. In *Proceedings of the 22th International Conference on Very Large Data Bases (VLDB’96)*. 402–412.
- [65] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numerSense: Probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 6862–6868. <https://doi.org/10.18653/v1/2020.emnlp-main.557>
- [66] Michael Luggen, Julien Audiffren, Djellel Difallah, and Philippe Cudré-Mauroux. 2021. Wiki2Prop: A multimodal approach for predicting Wikidata properties from Wikipedia. In *Proceedings of the Web Conference 2021 (WWW’21)*. Association for Computing Machinery, 2357–2366. <https://doi.org/10.1145/3442381.3450082>

- [67] Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. 2019. Non-parametric class completeness estimators for collaborative knowledge graphs—The case of Wikidata. In *The Semantic Web—ISWC 2019: 18th International Semantic Web Conference*. Springer, 453–469.
- [68] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*. Association for Computational Linguistics, 9802–9822. <https://doi.org/10.18653/v1/2023.acl-long.546>
- [69] Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*. Association for Computational Linguistics, 523–534.
- [70] MediaWiki. 2022. ORES. Retrieved 1-July-2022 from <https://www.mediawiki.org/w/index.php?title=ORES>
- [71] Paramita Mirza, Simon Razniewski, Fariz Darari, and Gerhard Weikum. 2018. Enriching knowledge bases with counting quantifiers. In *International Semantic Web Conference (ISWC)*. Springer, 179–197.
- [72] Paramita Mirza, Simon Razniewski, and Werner Nutt. 2016. Expanding Wikidata’s parenthood information by 178%, or how to mine relation cardinality information. In *The Semantic Web: ISWC 2016 Posters & Demonstrations Trac*. CEUR-WS.org.
- [73] Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. 2017. Domain-targeted, high precision knowledge extraction. 5 (2017), 233–246. https://doi.org/10.1162/tacl_a_00058
- [74] Amihai Motro. 1989. Integrity= validity+completeness. 14, 4 (1989), 480–502. <https://doi.org/10.1145/76902.76904>
- [75] Emir Muñoz, Pasquale Minervini, and Matthias Nickles. 2019. Embedding cardinality constraints in neural link predictors. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC'19)*. Association for Computing Machinery, 2243–2250. <https://doi.org/10.1145/3297280.3297502>
- [76] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*. Association for Computational Linguistics, 1135–1145.
- [77] Tuan-Phong Nguyen and Simon Razniewski. 2022. Materialized knowledge bases from commonsense transformers. (2022), 36–42. <https://doi.org/10.18653/v1/2022.csr-1.5>
- [78] Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. 2023. Refined commonsense knowledge from large-scale web contents. *Trans. Knowl. Datab. Eng.* 35, 8 (2023), 8431–8447. <https://doi.org/10.1109/TKDE.2022.3206505>
- [79] Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *Proceedings of the Web Conference 2021 (WWW'21)*. Association for Computing Machinery, 2636–2647. <https://doi.org/10.1145/3442381.3449827>
- [80] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale knowledge graphs: Lessons and challenges: five diverse technology companies show how it’s done. *Queue* 17, 2 (2019), 48–75. <https://doi.org/10.1145/3329781.3332266>
- [81] Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2022. Unik-QA: Unified representations of structured and unstructured knowledge for open-domain question answering. (2022), 1535–1546. <https://doi.org/10.18653/v1/2022.findings-naacl.115>
- [82] OpenAI. 2022. Introducing ChatGPT. Retrieved from <https://openai.com/blog/chatgpt>
- [83] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeljanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. *Large Language Models and Knowledge Graphs: Opportunities and Challenges*. arXiv:2308.06374
- [84] Pierre-Henri Paris, Syrine El Aoud, and Fabian M. Suchanek. 2021. The vagueness of vagueness in noun phrases. In *International Conference on Automated Knowledge Base Construction (AKBC'21)*. <https://imt.hal.science/hal-03344675>
- [85] Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semant.Web* 8, 3 (2017), 489–508. <https://doi.org/10.3233/SW-160218>
- [86] Thomas Pellissier Tanon, Daria Stepanova, Simon Razniewski, Paramita Mirza, and Gerhard Weikum. 2017. Completeness-aware rule learning from knowledge graphs. In *The Semantic Web – ISWC 2017*.
- [87] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. *Language Models as Knowledge Bases?* Association for Computational Linguistics, Hong Kong, 2463–2473. <https://doi.org/10.18653/v1/D19-1250>

- [88] Marco Ponzá, Paolo Ferragina, and Soumen Chakrabarti. 2017. A two-stage framework for computing entity relatedness in Wikipedia. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM'17)*. Association for Computing Machinery, 1867–1876. <https://doi.org/10.1145/3132847.3132890>
- [89] Soumajit Pramanik, Jesujoba Alabi, Rishiraj Saha Roy, and Gerhard Weikum. 2021. UNIQORN: Unified question answering over RDF knowledge graphs and natural language text. arXiv:2108.08614.
- [90] Radityo Eko Prasajo, Fariz Darari, Simon Razniewski, and Werner Nutt. 2016. Managing and consuming completeness information for Wikidata using COOL-WD. *CEUR Workshop Proceedings on 7th International Workshop on Consuming Linked Data 1666* (2016).
- [91] Péter Rácz. 2013. *Salience in Sociolinguistics: A Quantitative Approach*. Vol. 84. Walter de Gruyter.
- [92] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.
- [93] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 784–789. <https://doi.org/10.18653/v1/P18-2124>
- [94] Simon Razniewski. 2016. Optimizing update frequencies for decaying information. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM'16)*. Association for Computing Machinery, 1191–1200. <https://doi.org/10.1145/2983323.2983719>
- [95] Simon Razniewski, Vevake Balaraman, and Werner Nutt. 2017. Doctoral advisor or medical condition: Towards entity-specific rankings of knowledge base properties. In *Advanced Data Mining and Applications: 13th International Conference (ADMA'17)*. Springer, 526–540.
- [96] Simon Razniewski and Priyanka Das. 2020. Structured knowledge: Have we made progress? An extrinsic study of KB coverage over 19 years. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM'20)*. Association for Computing Machinery, 3317–3320. <https://doi.org/10.1145/3340531.3417447>
- [97] Simon Razniewski, Nitisha Jain, Paramita Mirza, and Gerhard Weikum. 2019. Coverage of information extraction from sentences and paragraphs. 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)*. Association for Computational Linguistics, Hong Kong, 5771–5776. <https://doi.org/10.18653/v1/D19-1583>
- [98] Simon Razniewski, Flip Korn, Werner Nutt, and Divesh Srivastava. 2015. Identifying the extent of completeness of query answers over partially complete databases. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD'15)*. Association for Computing Machinery, 561–576. <https://doi.org/10.1145/2723372.2750544>
- [99] Simon Razniewski and Werner Nutt. 2011. Completeness of queries over incomplete databases. *Proc. VLDB Endow.* 4, 11 (2011), 749–760. <https://doi.org/10.14778/3402707.3402715>
- [100] Simon Razniewski, Fabian M. Suchanek, and Werner Nutt. 2016. But what do we actually know? In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC)*. 40–44.
- [101] Nils Reimers and Iryna Gurevych. 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Association for Computational Linguistics, Hong Kong, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- [102] Raymond Reiter. 1981. On closed world data bases. In *Readings in Artificial Intelligence*, Elsevier, 119–140.
- [103] Daniel Ringler and Heiko Paulheim. 2017. One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co. In *Advances in Artificial Intelligence: 40th Annual German Conference on AI, Proceedings 40 (KI'17)*. Springer, 366–372.
- [104] D. S. Robson and H. A. Regier. 1964. Sample size in Petersen mark–recapture experiments. *Trans. Am. Fisher. Societ.* 3 (1964), 215–226.
- [105] Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhdeo, and Gerhard Weikum. 2019. Commonsense properties from query logs and question answering forums. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM'19)*. Association for Computing Machinery, 1411–1420. <https://doi.org/10.1145/3357384.3357955>
- [106] Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Trans. Assoc. Computat. Ling.* 3 (2015), 1–13. DOI: https://doi.org/10.1162/tacl_a_00118
- [107] Tara Safavi and Danai Koutra. 2020. Generating negative commonsense knowledge. In *4th Knowledge Representation and Reasoning Meets Machine Learning Workshop at NeurIPS*.
- [108] Tara Safavi, Jing Zhu, and Danai Koutra. 2021. NegatER: Unsupervised discovery of negatives in commonsense knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 5633–5646. <https://doi.org/10.18653/v1/2021.emnlp-main.456>
- [109] Ognjen Savković, Paramita Mirza, Alex Tomasi, and Werner Nutt. 2013. Complete approximations of incomplete queries. *Proc. VLDB Endow.* 6, 12 (2013), 1378–1381. DOI: <https://doi.org/10.14778/2536274.2536320>

- [110] Amit Singhal. 2012. Introducing the Knowledge Graph: Things, not strings. Retrieved from <https://blog.google/products/search/introducing-knowledge-graph-things-not/>
- [111] Sneha Singhania, Jan-Christoph Kalo, Simon Razniewski, and Jeff Z. Pan. 2023. Knowledge base construction from pre-trained language models. In *Challenge at ISWC 2023*.
- [112] Sneha Singhania, Tuan-Phong Nguyen, and Simon Razniewski. 2022. LM-KBC: Knowledge base construction from pre-trained language models. In *Challenge at ISWC 2022*. Retrieved from <https://ceur-ws.org/Vol-3274/paper1.pdf>
- [113] Sneha Singhania, Simon Razniewski, and Gerhard Weikum. 2023. Extracting multi-valued relations from language models. In *Workshop on Representation Learning for NLP (RepL4NLP 2023)*.
- [114] Arnaud Soulet, Arnaud Giacometti, Béatrice Markhoff, and Fabian M. Suchanek. 2018. Representativeness of knowledge bases with the generalized Benford's law. In *The Semantic Web—ISWC 2018: 17th International Semantic Web Conference*. Springer, 374–390.
- [115] Diana Sousa, Andre Lamurias, and Francisco M. Couto. 2020. Improving accessibility and distinction between negative results in biomedical relation extraction. *Genom. Inform.* (2020).
- [116] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, 4444–4451.
- [117] Jannik Strötgen and Michael Gertz. 2010. HeideITime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*. Association for Computational Linguistics, 321–324.
- [118] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. Association for Computing Machinery, 697–706. <https://doi.org/10.1145/1242572.1242667>
- [119] Fabian M. Suchanek, Jonathan Lajus, Armand Boschin, and Gerhard Weikum. 2019. Knowledge representation and rule mining in entity-centric knowledge bases. In *Reasoning Web. Explainable Artificial Intelligence: 15th International Summer School 2019, Tutorial Lectures*, 110–152.
- [120] Fabian M. Suchanek and Anh Tuan Luu. 2023. Knowledge bases and language models: Complementing forces. In *RuleML+RR Invited Paper*.
- [121] Shelley E. Taylor and Susan T. Fiske. 1978. Saliency, attention, and attribution: Top of the head phenomena. In *Advances in Experimental Social Psychology*, Vol. 11.
- [122] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
- [123] Beth Trushkowsky, Tim Kraska, Michael J. Franklin, and Purnamrita Sarkar. 2013. Crowdsourced enumeration queries. In *IEEE 29th International Conference on Data Engineering (ICDE'13)*. IEEE, 673–684.
- [124] Blerta Veseli, Sneha Singhania, Simon Razniewski, and Gerhard Weikum. 2023. Evaluating language models for knowledge base completion. In *European Semantic Web Conference (ESWC)*. Springer, 227–243.
- [125] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Commun. ACM* 57, 10 (2014), 78–85. <https://doi.org/10.1145/2629489>
- [126] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adaptor: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021)*. 1405–1418.
- [127] Gerhard Weikum, Luna Dong, Simon Razniewski, and Fabian M. Suchanek. 2021. Machine knowledge: Creation and curation of comprehensive knowledge bases. *Foundations and TrendsR in Databases* 10, 2–4 (2021), 108–490.
- [128] Ian H. Witten and David N. Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*. AAAI Press, 25–30.
- [129] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*. Association for Computational Linguistics, 23–30. <https://doi.org/10.18653/v1/2020.emnlp-demos.4>
- [130] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. 2016. Quality assessment for linked data: A survey. *Semantic Web* 7, 1 (2016), 63–93.
- [131] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. *Explainability for Large Language Models: A Survey*. arXiv preprint arXiv:2309.01029 (2023).

Received 30 November 2022; revised 9 November 2023; accepted 21 December 2023