Semantic Web 13 (2022) 925–966
DOI 10.3233/SW-212865
IOS Press

Survey on English Entity Linking on Wikidata: Datasets and approaches

Cedric Möller [a,*], Jens Lehmann [b,c] and Ricardo Usbeck [a,d]

[a] *Semantic Systems Group, Universität Hamburg, Mittelweg 177, 20148 Hamburg, Germany*
E-mails: cedric.moeller@uni-hamburg.de, ricardo.usbeck@uni-hamburg.de
[b] *NetMedia Department, Fraunhofer IAIS, Zwickauer Straße 46, 01069 Dresden, Germany*
E-mail: jens.lehmann@iais.fraunhofer.de
[c] *University of Bonn, Endenicher Allee 19a, 53115, Bonn, Germany*
E-mail: jens.lehmann@cs.uni-bonn.de
[d] *HITeC Hamburg e.V., Vogt-Kölln-Straße 30, 22527 Hamburg, Germany*

Editors: Julia Bosque-Gil, University of Zaragoza, Spain; Milan Dojchinovski, Czech Technical University in Prague, Czech Republic; Philipp Cimiano, Bielefeld University, Germany
Solicited reviews: Vasilis Efthymiou, Institute of Computer Science-FORTH, Greece; Albert Weichselbraun, Chur University of Applied Sciences, Switzerland; Filip Ilievski, University of Southern California, USA; one anonymous reviewer

Abstract. Wikidata is a frequently updated, community-driven, and multilingual knowledge graph. Hence, Wikidata is an attractive basis for Entity Linking, which is evident by the recent increase in published papers. This survey focuses on four subjects: (1) Which Wikidata Entity Linking datasets exist, how widely used are they and how are they constructed? (2) Do the characteristics of Wikidata matter for the design of Entity Linking datasets and if so, how? (3) How do current Entity Linking approaches exploit the specific characteristics of Wikidata? (4) Which Wikidata characteristics are unexploited by existing Entity Linking approaches? This survey reveals that current Wikidata-specific Entity Linking datasets do not differ in their annotation scheme from schemes for other knowledge graphs like DBpedia. Thus, the potential for multilingual and time-dependent datasets, naturally suited for Wikidata, is not lifted. Furthermore, we show that most Entity Linking approaches use Wikidata in the same way as any other knowledge graph missing the chance to leverage Wikidata-specific characteristics to increase quality. Almost all approaches employ specific properties like labels and sometimes descriptions but ignore characteristics such as the hyper-relational structure. Hence, there is still room for improvement, for example, by including hyper-relational graph embeddings or type information. Many approaches also include information from Wikipedia, which is easily combinable with Wikidata and provides valuable textual information, which Wikidata lacks.

Keywords: Entity Linking, Entity Disambiguation, Wikidata

1. Introduction

1.1. Motivation

Entity Linking (EL) is the task of connecting already marked mentions in an utterance to their corresponding entities in a knowledge graph (KG), see Fig. 1. In the past, this task was tackled by using popular knowledge

*Corresponding author. E-mail: cedric.moeller@uni-hamburg.de.

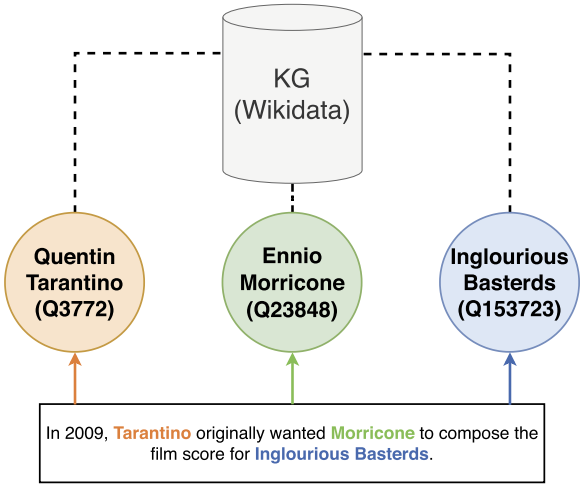


Fig. 1. Entity linking – mentions in the text are linked to the corresponding entities (color-coded) in a knowledge graph (here: Wikidata).

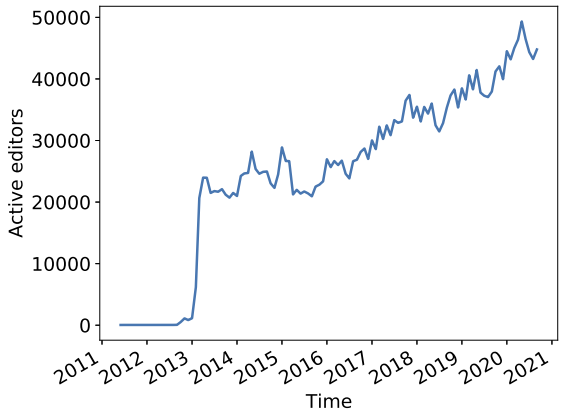


Fig. 2. Active editors in Wikidata [36].

bases such as DBpedia [67], Freebase [12] or Wikipedia. While the popularity of those is still imminent, another alternative, named Wikidata [120], appeared.

Wikidata follows a similar philosophy as Wikipedia as it is curated by a continuously increasing community, see Fig. 2. However, Wikidata differs in the way knowledge is stored – information is stored in a structured format via a knowledge graph (KG). An important characteristic of Wikidata is its inherent multilingualism. While Wikipedia articles exist in multiple languages, Wikidata information are stored using language-agnostic identifiers. This is of advantage for multilingual entity linking. DBpedia, Freebase or Yago4 [109] are KGs too which can become outdated over time [93]. They rely on information extracted from other sources in contrast to the Wikidata knowledge which is inserted by a community. Given an active community, this leads to Wikidata being frequently and timely updated – another characteristic. Note that DBpedia also stays up-to-date but has a delay of a month[1] while Wikidata dumps are updated multiple times a month. There are up-to-date services to access knowledge for both KGs, Wikidata and DBpedia (cf. DBpedia Live[2]), but full dumps are preferred as else the FAIR replication [126]

[1] https://release-dashboard.dbpedia.org/
[2] https://wiki.dbpedia.org/online-access/DBpediaLive

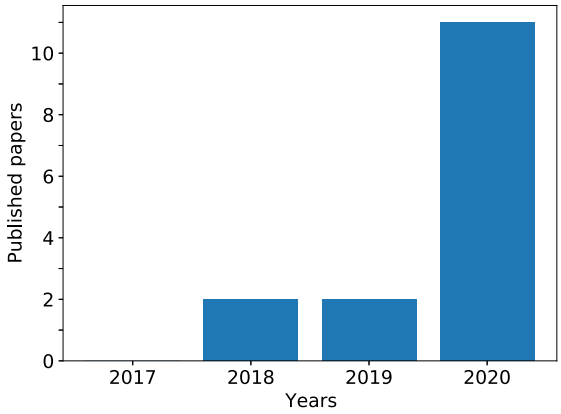


Fig. 3. Publishing years of included Wikidata EL papers (Table 11).

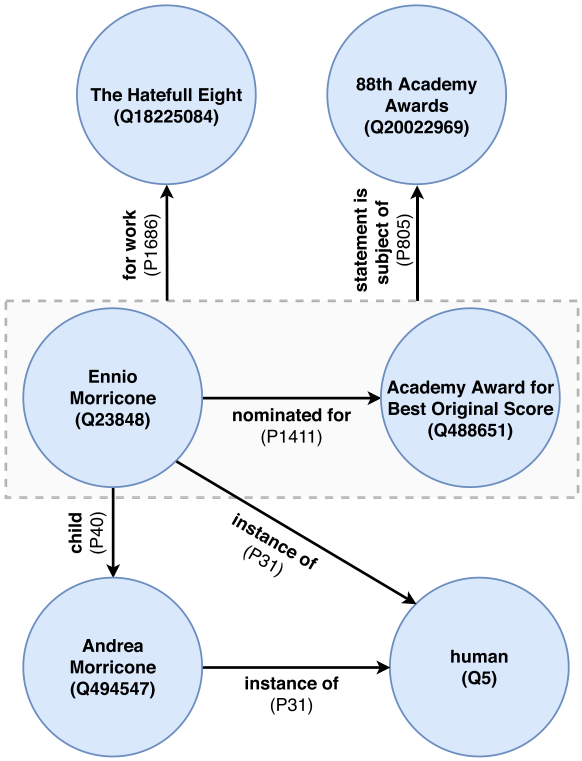


Fig. 4. Wikidata subgraph – dashed rectangle represents a claim with attached qualifiers.

of research results based on the KG is hindered. Another Wikidata characteristic interesting for Entity Linkers, are hyper-relations (see Fig. 4 for an example graph), which might affect their abilities and performance.

Therefore, it is of interest how existing approaches incorporate these characteristics. However, existing literature lacks an exhaustive analysis which examines Entity Linking approaches in the context of Wikidata.

Ultimately, this survey strives to expose the benefits and associated challenges which arise from the use of Wikidata as the target KG for EL. Additionally, the survey provides a concise overview of existing EL approaches, which is essential to (1) avoid duplicated research in the future and (2) enable a smoother entry into the field of

Wikidata EL. Similarly, we structure the dataset landscape which helps researchers find the correct dataset for their EL problem.

The focus of this survey lies on EL approaches, which operate on already marked mentions of entities, as the task of Entity Recognition (ER) is much less dependent on the characteristics of a KG. However, due to the recent uptake of research on EL on Wikidata, there is only a low number of EL-only publications. To broaden the survey's scope, we also consider methods that include the task of ER. We do not restrict ourselves regarding the type of models used by the entity linkers.

This survey limits itself to all EL approaches supporting the English language as most frequent language, and thus, a better comparison of the approaches and datasets is possible. We also include approaches that support multiple languages. The existence of such approaches for Wikidata is not surprising as an important characteristic of Wikidata is the support of a multitude of languages.

1.2. Research questions and contributions

First, we want to develop an overview of datasets for EL on Wikidata. Our survey analyses datasets and whether they are designed with Wikidata in mind and if so, in what way? Thus, we post the following two research questions:

RQ 1: Which Wikidata EL datasets exist, how widely used are they and how are they constructed?

RQ 2: Do the characteristics of Wikidata matter for the design of EL datasets and if so, how?

To answer those two research questions, an overview of the structure of Wikidata and the amount of information it contains (see Section 4) is given. All current Wikidata-specific EL datasets are gathered and analyzed with the research questions in mind. Furthermore, we discuss how the characteristics of Wikidata might affect the design of datasets (see Section 5).

EL approaches use many kinds of information like labels, popularity measures, graph structures, and more. This multitude of possible signals raises the question of how the characteristics of Wikidata are used by the current state of the art of EL on Wikidata. Thus, the third research question is:

RQ 3: How do current Entity Linking approaches exploit the specific characteristics of Wikidata?

In particular, which Wikidata-specific characteristics contribute to the solution? Wikidata-specific characteristics mean characteristics that are part of Wikidata but not necessarily only occurring in Wikidata.

Lastly, we identify what kind of characteristics of Wikidata are of importance for EL but are insufficiently considered. This raises the last research question:

RQ 4: Which Wikidata characteristics are unexploited by existing Entity Linking approaches?

The last two questions are answered by gathering all existing approaches working on Wikidata systematically, analyzing them, and discussing the potential and challenges of Wikidata for EL (see Section 6).

This survey makes the following contributions:

- An overview of all currently available EL datasets focusing on Wikidata
- An overview of all currently available EL approaches linking on Wikidata
- An analysis of the approaches and datasets with a focus on Wikidata characteristics
- A concise list of future research avenues

2. Survey methodology

There exist several different ways in which a survey can contribute to the research field [57]:

1. Providing an overview of current prominent areas of research in a field
2. Identification of open problems
3. Providing a novel approach tackling the extracted open problems (in combination with the identification of open problems)

Table 1
Qualifying and disqualifying criteria for approaches. "Semi-structured" in this table means that the entity mentions do not occur in natural language utterances but in more structured documents such as tables

Criteria	
Must satisfy all	Must not satisfy any
– Approaches that consider the problem of unstructured EL over Knowledge Graphs	– Approaches conducting Semi-structured EL
– Approaches where the target Knowledge Graph is Wikidata	– Approaches not doing EL in the English language

We analyse different recent and older surveys on EL and highlight specific areas which are not covered as well as our survey's novelties (see also Section 8). While some very recent surveys exist [2,81,101], they do not consider the different underlying Knowledge Graphs as a significant factor affecting the performance of EL approaches. Furthermore, barely any approaches included in other surveys are working on Wikidata and take the particular characteristics of Wikidata into account (see Section 7). Our survey fills these gaps by contributing according to Items 1 and 2.

Until December 18, 2020, we continuously searched for existing and newly released scientific work suitable for the survey. Note, this survey includes only scientific articles that were accessible to the authors.[3]

2.1. Approaches

Our selection of approaches stems from a search over the following search engines:

- Google Scholar
- Springer Link
- Science Direct
- IEEE Xplore Digital Library
- ACM Digital Library

To gather a wide choice of approaches, the following steps were applied. `Entity Linking`, `Entity Disambiguation` or variations of the phrases[4] had to occur in the title of the paper. The publishing year was not a criterion due to the small number of valid papers and the relatively recent existence of Wikidata. Any approach where `Wikidata` was not occurring once in the full text was not considered. The systematic search process resulted in exactly 150 papers and theses (including duplicates).

Following this search, the resulting papers were filtered again using the qualifying and disqualifying criteria which can be found in Table 1. This resulted in 15 papers and one master thesis in the end.

The search resulted in papers in the period from 2018 to 2020. While there exist EL approaches from 2016 [4,107] working on Wikidata, they did not qualify according to the criteria above.

2.2. Datasets

The dataset search was conducted in two ways. First, a search for potential datasets was performed via the same search engines as used for the approaches. Second, all datasets occurring in the system papers were considered if they fulfilled the criteria. The criteria for the inclusion of a dataset can be found in Table 2.

We filtered the dataset papers in the following way. First, in the title, `Entity Linking` or `Entity Disambiguation` or variations thereof had to occur, similar to the search for the Entity Linking approaches. Additionally, `dataset`, `data`, `corpus` or `benchmark` had to occur once in title[5] must occur in the title and `Wikidata` has

[3] https://www.projekt-deal.de/about-deal/

[4] Google Scholar search query: `(intitle:"entity" OR intitle:"entities") AND (intitle:"link" OR intitle:"linking" OR intitle:"disambiguate" OR entitle:"disambiguation") AND intext:"Wikidata"`.

[5] Google Scholar Search Query: `intext:"Wikidata" AND (intitle:dataset OR intitle:data OR intitle:benchmark OR intitle:corpus) AND (intitle:entity OR intitle:entities) AND (intitle:link OR intitle:linking OR intitle:disambiguate OR intitle:disambiguation)`.

Table 2
Qualifying and disqualifying criteria for the dataset search

Criteria	
Must satisfy all	Must not satisfy any
– Datasets that are designed for EL or are used for evaluation of Wikidata EL	– Datasets without English utterances
– Datasets must include Wikidata identifiers from the start; an existing dataset later mapped to Wikidata is not permitted	

to appear at least once in the full text. Due to those keywords, other datasets suitable for EL, but constructed for a different purpose like KG population, were not included. This resulted in 26 papers (including duplicates). Of those, only two included Wikidata identifiers and focused on English.

Eighteen datasets were accompanying the different approaches. Many of those did not include Wikidata identifiers from the start. This made them less optimal for the examination of the influence of Wikidata on the design of datasets. They were included in the section about the approaches but not in the section about the Wikidata datasets.

After the removal of duplicates, 11 Wikidata datasets were included in the end.

3. Problem definition

EL is the task of linking an entity mention in unstructured or semi-structured data to the correct entity in a KG. The focus of this survey lies in unstructured data, namely, natural language utterances.

3.1. General terms

Utterance An utterance u is defined as a sequence of n words w.

$$u = (w_0, w_1, \ldots w_{n-1})$$

Entity There exists no universally agreed-on definition of an entity in the context of EL [97]. According to the Oxford Dictionary, an entity is:

> "something that exists separately from other things and has its own identity" [82]

What elements of a KG correspond to entities depends on the KG itself. In the case of Wikidata, we define it as follows:

> Any Wikidata item is an entity.

In Section 4, we further define Wikidata items. Many EL approaches limit the space of valid entities. Usually, named entities like a specific person (e.g. `Barack Obama`), an organization (e.g. `NASA`) or a movie (e.g. `The Hateful Eight`) are desirable to link. In general, any entity which can be denoted with a proper noun is a named entity. But sometimes, also common entities like concepts (e.g. `dog` or `theater`) are included. What exactly is linked depends on the use case [97].

Knowledge graph While the term knowledge graph was already used before, the popularity increased drastically after Google introduced the `Knowledge Graph` in 2012 [28,103]. However, similar to an entity, there exists no unanimous definition of a KG [28,52]. For example, Färber et al. define a KG as an RDF graph [35]. However, a KG being an RDF graph is a strict assumption. While the Wikidata graph is available in the RDF format, the main output format is JSON. Freebase, often called a KG, did not provide the RDF format until a year after its launch [11]. Paulheim defines it less formal as:

> "A knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains." [84]

But constraint (iv) of Paulheims definition alienates commercial KGs, focusing on a single domain, such as a financial, medical, or geographical one. As no unanimously agreed definition exists, we define a knowledge graph very broadly in the following and the Wikidata KG more concrete in Section 4. The term KG is often used as a synonym for the term knowledge base (KB), but they are not the same [54]. No single definition for a KB exists either. Jarke et al. define it as "a representation of heuristic and factual information, often in the form of facts, assertions and deduction rules" [54]. The term is often loosely used to describe a system that is able to store knowledge in the form of structured or unstructured information. While any KG is a KB, not any KB is a KG. The main difference is that a KB does not have to be graph-structured.

In this survey, a knowledge graph is defined as a directed graph $G = (V, E, \mathcal{R})$ consisting of vertices V, edges E and relations $\mathcal{R}$. A subset of the vertices corresponds to entities $\mathcal{E}$ or literals $\mathcal{L}$. A literal is a concrete value of information like the height or a name of an entity. A literal vertex has incoming edges but no outgoing ones. Other types of vertices might exist depending on the KG. E is a set $\{e_1, \ldots, e_{|E|}\}$ of edges with $e_j \in V \times \mathcal{R} \times V$ where relations $\mathcal{R}$ assign a certain meaning to the connection between entities. Such edges are also called triples. There are special subtypes of KGs, e.g., hyper-relational graphs such as Wikidata.

Hyper-relational knowledge graphs In a hyper-relational knowledge graph, statements can be specified by more information than a single relation. Multiple relations are, therefore, part of a statement. In case of a hyper-relational graph $\mathcal{G} = (V, E, \mathcal{R})$, E is a list $(e_1, \ldots, e_n)$ of edges with $e_j \in V \times \mathcal{R} \times V \times \mathcal{P}(\mathcal{R} \times V)$ for $1 \leqslant j \leqslant n$, where $\mathcal{P}$ denotes the power set. A hyper-relational fact $e_j \in E$ is usually written as a tuple $(s, r, o, \mathcal{Q})$, where $\mathcal{Q}$ is the set of *qualifier pairs* $\{(qr_i, qv_i)\}$ with *qualifier relations* $qr_i \in \mathcal{R}$ and *qualifier values* $qv_i \in V$. The triple (s, r, o) is referred to as the *main triple* of the fact. $\mathcal{Q}_j$ denotes the qualifier pairs of e_j [37]. For example, the `nominated for` edge in Fig. 4 has two additional qualifier relations and would be represented as `(Ennio Morricone, nominated for, Academy Award for Best Original Score, {(for work, The Hateful Eight), (statement is subject of, 88th A-cademy Awards)})`.

3.2. Tasks

Since not only approaches that solely do EL were included in the survey, Entity Recognition will also be defined.

Entity recognition ER is the task of identifying the mention span

$$m = (w_i, \ldots, w_k) | 0 \leqslant i \leqslant k \leqslant n - 1$$

of all entities in an utterance u. Each such span is called an entity mention m. The word or word sequence referring to an entity is also known as the surface form of an entity. An utterance can contain more than one entity, often also consisting of more than one word. Sometimes, a broad type of an entity is classified too. Usually, those are `person`, `location` and `organization`. Some of the considered approaches do such a classification task and also use it to improve the EL.

It is also up to debate what an entity mention is. In general, a literal reference to an entity is considered a mention. But whether to include pronouns or how to handle overlapping mentions depends on the use case.

Entity linking The goal of EL is to find a mapping function that maps all found mentions to the correct KG entities and also to identify if an entity mention does not exist in the KG.

In general, EL takes the utterance u and all k identified entity mentions $M = (m_1, \ldots m_k)$ in the utterance and links each of them to an element of the set $(\mathcal{E} \cup \{NIL\})$. The *NIL* element is added to the set of vertices to be able to signalize that the entity, that the mention is referring to, is not known to the KG. Such a *NIL* entity is also called an out-of-KG entity. Another way to handle such unknown entities is to create emerging entities [50]. In that case, the entity is still unknown to the KG, but after encountering it, it is separately stored using information like the provided entity mentions. Now no single *NIL* entity, but a growing set of emerging entities exists. EL is then done using the entities in the KG and all already encountered emerging entities. While all KG-unknown entities point to the same single *NIL* entity, they might point to different emerging entities.

EL is often split into two subtasks. First, potential candidates for an entity are retrieved from a KG. This is necessary as doing EL over the whole set of entities is often intractable. This *Candidate generation* is usually performed via efficient metrics measuring the similarities between mentions in the utterance and entities in the KG. The result is a set of candidates $C = \{c_0, \ldots, c_l\}$ for each entity mention m in the utterance. After limiting the space of possible entities, one of the available candidates is chosen for each entity. This is done via a *candidate ranking* algorithm, which assigns a rank to each candidate. The assignment is done by computing a score for each candidate signalizing how likely it is the correct entity. The candidate with the highest score is chosen as the correct entity for the mention.

There are two different categories of reranking methods are called *local* or *global* [91].

$$score_{local} : C \times M \to \mathbb{R}$$
$$\text{given by } (c, m) \mapsto score_{local}(c, m)$$

where $score_{local}$ is a local scoring function of a candidate. The goal is then to optimize the objective function:

$$A^* = \arg\max_A \sum_{i=1}^{k} score_{local}(a_i, m_i) | a_i \in C_i$$

where $A = \{a_1, \ldots, a_k\} \in \mathcal{P}(\mathcal{E})$ is an assignment of one candidate to each entity mention m_i. $\mathcal{P}(*)$ is the power set operator.

The rank assignment and score calculation of the candidates of one entity is often not independent of the other entities' candidates. In this case, the ranking will be done by including the whole assignment via a global scoring function:

$$score_{global} : \mathcal{P}(\mathcal{E}) \to \mathbb{R} \text{ given by } A \mapsto score_{global}(A)$$

The objective function is then:

$$A^* = \arg\max_A \left[\sum_{i=1}^{k} score_{local}(a_i, m_i) \right] + score_{global}(A) | a_i \in C_i$$

Note, there also exists some ambiguity in the objective of linking itself. For example, there exists a Wikidata entity `2014 FIFA World Cup` and an entity `FIFA World Cup`. There is no unanimous solution on how to link the entity mention in the utterance `In 2014, Germany won the FIFA World Cup`.

Sometimes EL is also called Entity Disambiguation, which we see more as part of EL, namely where entities are disambiguated via the candidate ranking.

There exist multiple special cases of EL. *Multilingual EL* tries to link entity mentions occurring in utterances of different languages to one shared KG, for example, English, Spanish or Chinese utterances to one language-agnostic KG. Formally, an entity mention m in some utterance u of some context language l_c has to be linked to a language-agnostic KG which includes information in multiple languages $L_{KG} = \{l_1, \ldots, l_k\}$ where l_c can but has not to be an element of L_{KG} [16].

Cross-lingual EL tries to link entity mentions in utterances in different languages to a KG in one dedicated language, for example, Spanish and German utterances to an English KG [92]. In that case, the multilingual EL problem gets constrained to $L_{KG} = \{l_{KG}\}$ where $l_c \neq l_{KG}$.

In *zero-shot EL*, the entities during test time $\mathcal{E}_{test}$ are not available at training time $\mathcal{E}_{train}$.

$$\mathcal{E}_{test} \cap \mathcal{E}_{train} = \emptyset \quad \text{where } \mathcal{E}_{test} \subset \mathcal{E}, \mathcal{E}_{train} \subset \mathcal{E}$$

Thus, the entity linker must be able to handle unseen entities. The term was coined by Logeswaran et al. [73], but they limited the task to only have descriptions available while our definition does not include such a limitation.

Table 3
KG statistics by [109]

KG	#Entities in million	#Labels/Aliases in million	Last updated
Wikidata	78	442	Up to 4 times a month*
DBpedia	5	22	Monthly
Yago4	67	371	November 2019

*https://dumps.wikimedia.org/wikidatawiki/entities/

KB/KG-agnostic EL approaches are able to support different KBs respectively KGs, often multiple in parallel. For example, a KG must be available in RDF format. We refer the interested reader to central works [76,114,137] or our Appendix.

4. Wikidata

Wikidata is a community-driven knowledge graph edited by humans and machines. The Wikidata community can enrich the content of Wikidata by, for example, adding/changing/removing entities, statements about them, and even the underlying ontology information. As of July 2020, it contained around 87 million items of structured data about various domains. Seventy-three million items can be interpreted as entities due to the existence of an `is instance` property. As a comparison, DBpedia contains around 5 million entities [109]. Note that the `is instance` property includes a much broader scope of entities than the ones interpreted as entities for DBpedia. In comparison to other similar KGs, the Wikidata dumps are updated most frequently (Table 3). But note that this only applies to the dumps, if one considers direct access via the Website or a SPARQL endpoint, both, Wikidata[6,7] and DBpedia[8,9] provide continuously updated knowledge.

4.1. Definition

Wikidata is a collection of *entities* where each such entity has a page on Wikidata. An entity can be either an *item* or a *property*. Note, an entity in the sense of Wikidata is generally not the same as an entity one links to via EL. For example, Wikidata entities are also properties that describe relations between different items. Linking to such relations is closer to Relation Extraction [9,70,104]. Furthermore, many items are more abstract classes, which are usually not considered as entities linked to in EL. Note that if not mentioned otherwise, if we speak about entities, entities in the context of EL are meant.

Item Topics, classes, or objects are defined as items. An item is enriched with more information using statements about the item itself. In general, items consist of one label, one description, and aliases in different languages. An unique and language-agnostic identifier identifies items in the form `Q[0-9]+`. An example of an item can be found in Fig. 5.

For example, the item with the identifier `Q23848` has the label `Ennio Morricone`, two aliases, `Dan Savio` and `Leo Nichols`, and `Italian composer, orchestrator and conductor (1928-2020)` as description at the point of writing. The corresponding Wikidata page can also be seen in Fig. 5.

Property A property specifies a relation between items or literals. Each property also has an identifier similar to an item, specified by `P[0-9]+`. For instance, a property `P19` specifies the place of birth `Rome` for `Ennio Morricone`. In NLP, the term `relation` is commonly used to refer to a connection between entities. A property in the sense of Wikidata is a type of relation. To not break with the terminology used in the examined papers, when we talk about relations, we always mean Wikidata properties if not mentioned otherwise.

[6]https://www.wikidata.org/
[7]https://query.wikidata.org
[8]https://www.dbpedia.org/resources/live/
[9]https://www.dbpedia.org/resources/live/dbpedia-live-sync/

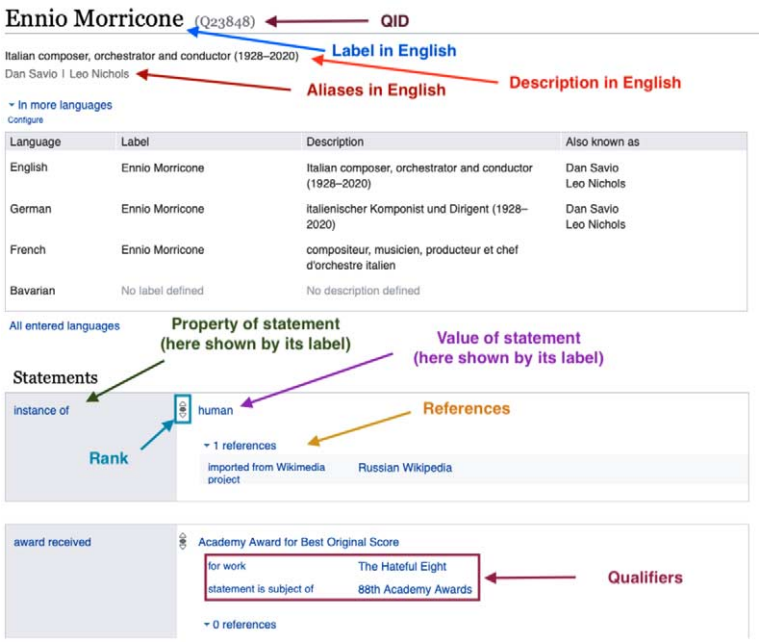


Fig. 5. Example of an item in Wikidata.

Statement A statement introduces information by giving structure to the data in the graph. It is specified by a *claim*, and *references*, *qualifiers* and *ranks* related to the claim. Statements are assigned to items in Wikidata. A claim is defined as a pair of property and some value. A value can be another item or some literal. Multiple values are possible for a property. Even an `unknown value` and a `no value` exists.

References point to sources making the claims inside the statements verifiable. In general, they consist of the source and date of retrieval of the claim.

Qualifiers define the value of a claim further by contextual information. For example, a qualifier could specify for how long one person was the spouse of another person. Qualifiers enable Wikidata to be hyper-relational (see Section 3.1). Structures similar to qualifiers also exist in some other knowledge graphs, such as the inactive Freebase in the form of Compound Value Types [12].

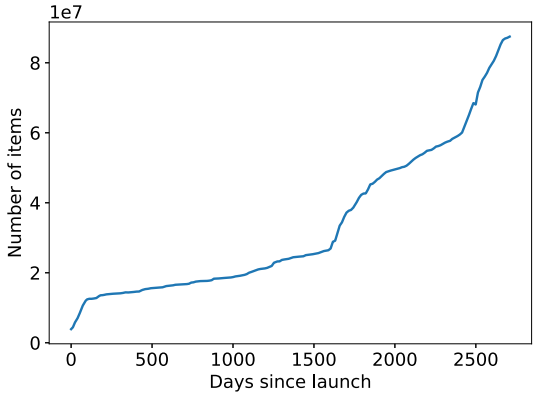
Ranks are used if multiple values are valid in a statement. If the population of a country is specified in a statement, it might also be useful to have the populations of past years available. The most up-to-date population information usually has then the highest rank and is thus usually the most desirable claim to use.

Statements can be also seen in Fig. 5 at the bottom. For example, it is defined that `Ennio Morricone` is an `instance of` the class `human`. This is also an example for the different types of items. While `Ennio Morricone` is an entity in our sense, `human` is a class.

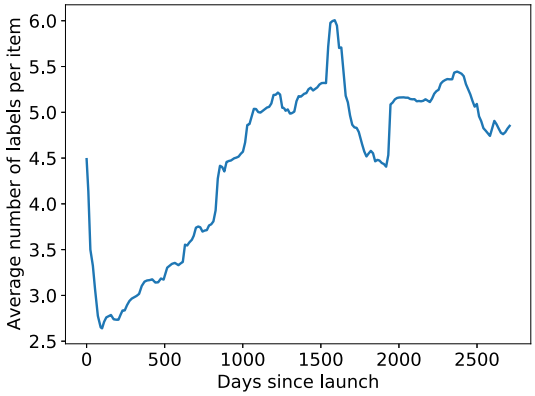
Other structural elements The aforementioned elements are essential for Wikidata, but more do exist. For example, there are entities (in the sense of Wikidata) corresponding to Lexemes, Forms, Senses or Schemas. Lexemes, Forms and Senses are concerned with lexicographical information, hence words, phrases and sentences themselves. This is in contrast to Wikidata items and properties, which are directly concerned with things, concepts and ideas. Schemas formally subscribe to subsets of Wikidata entities. For example, any Wikidata item which has `actor` as its `occupation` is an `instance of` the class `human`. Both, lexicographical and schema information, are usually not directly of relevance for EL. Therefore, we refrain from introducing them in more detail.

For more information on Wikidata, see the paper by Denny Vrandečić and Markus Krötzsch [120].

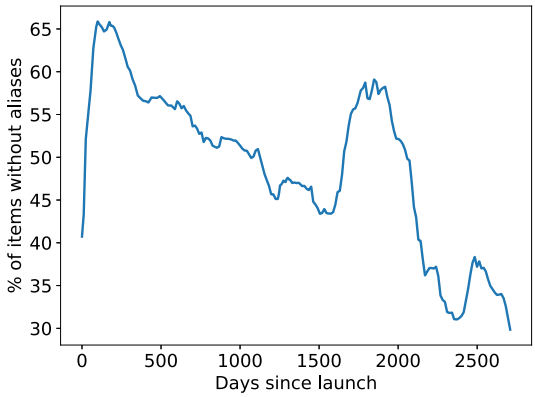
Differences in structure to other knowledge graphs DBpedia extracts its information from Wikipedia and Wikidata. It maps the information to its own ontology. DBpedia's statements consist of only single triples (`<subject, predicate, object>`) since it follows the RDF specification [67]. Additional information like qualifiers, ref-



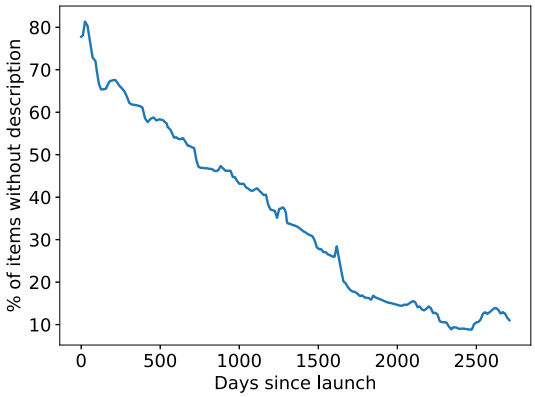
(a) Number of items of Wikidata since launch [73].



(b) Average number of labels (+ aliases) per item [73].



(c) Percentage of items without any aliases [73].



(d) Percentage of items without a description [73].

Fig. 6. Statistics on Wikidata based on [74].

erences or ranks do not exist. But it can be modeled via additional triples. As it is no inherent feature of DBpedia, it is harder to use as there a no strict conventions. Entities in DBpedia have human-readable identifiers, and there exist entities per language [67] with partly differing information. Hence, for a single concept or thing, multiple DBpedia entities might exist. For example, the English entity of the city Munich[10] has 25 entities as `dbo:administrativeDistrict` assigned. The German entity[11] only a single one. It seems that this originates from a different interpretation of the predicate `dbo:administrativeDistrict`.

Yago4 extracts all its knowledge from Wikidata but filters out information it deems inadequate. For example, if a property is used too seldom, it is removed. If a Wikidata entity does not have a class that exists in Schema.org,[12] it is removed. The RDF specification format is used. Qualifier information is included indirectly via separate triples. Rank information and references of statements do not exist. The identifiers follow either a human-readable form if available via Wikipedia or Wikidata or use the Wikidata QID. However, in contrast to DBpedia, only one entity exists per thing or concept [109].

For a thorough comparison of Wikidata and other KGs (in respect to Linked Data Quality [134]), please refer to the paper by Färber et al. [35].

[10] https://dbpedia.org/page/Munich
[11] http://de.dbpedia.org/page/Munchen
[12] https://schema.org

Table 4
Statistics – languages Wikidata (extracted from dump [125])

	Items	Properties
Number of languages	457	427
(average, median) of # languages per element (labels + descriptions)	29.04, 6	21.24, 13
(average, median) of # languages per element (labels)	5.59, 4	21.18, 6
(average, median) of # languages per element (descriptions)	26.10, 4	9.77, 6
% elements without English labels	15.41%	0%
% elements without English descriptions	26.23%	1.08%

4.2. Discussion

Novelties A useful characteristic of Wikidata is that the community can openly edit it. Another novelty is that there can be a plurality of facts, as contradictory facts based on different sources are allowed. Similarly, time-sensitive data can also be included by qualifiers and ranks. The population of a country, for example, changes from year to year, which can be represented easily in Wikidata. Lastly, due to their language-agnostic identifiers, Wikidata is inherently multilingual. Language only starts playing a role in the labels and descriptions of an item.

Strengths Due to the inclusion of information by the community, recent events will likely be included. The knowledge graph is thus much more up-to-date than most other KGs. Freebase is unsupported for years now, and DBpedia updates its dumps only every month. Note, the novel DBpedia live 2.0[13] is updated when changes to a Wikipedia page occur, but, as discussed, makes research harder to replicate. Thus, Wikidata is much more suitable and useful for industry applications such as smart assistants since it is the most complete open-accessible data source to date. In Fig. 6a, one can see that number of items in Wikidata is increasing steadily. The existence of labels and additional aliases (see Fig. 6b) helps EL as a too-small number of possible surface forms often lead to a failure in the candidate generation. DBpedia does, for example, not include aliases, only a single exact label;[14] to compensate, additional resources like Wikipedia are often used to extract a label dictionary of adequate size [76]. Even each property in Wikidata has a label [120]. Fully language model-based approaches are therefore more naturally usable [78]. Also, nearly all items have a description, see Fig. 6d. This short natural language phrase can be used for context similarity measures with the utterance. The inherent multilingual structure is intuitively useful for multilingual Entity Linking. Table 4 shows information about the use of different languages in Wikidata. As can be seen, item labels/aliases are available in up to 457 languages. But not all items have labels in all languages. On average, labels, aliases and descriptions are available in 29.04 different languages. However, the median is only 6 languages. Many entities will, therefore, certainly not have information in many languages. The most dominant language is English, but not all elements have label/alias/description information in English. For less dominant languages, this is even more severe. German labels exist, for example, only for 14%, and Samoan labels for 0.3%. Context information in the form of descriptions is also given in multiple languages. Still, many languages are again not covered for each entity (as can be seen by a median of only 4 descriptions per element). While the multilingual label and description information of items might be useful for language model-based variants, the same information for properties enables multilingual language models. Because, on average, 21.18 different languages are available per property for labels, one could train multilingual models on the concatenations of the labels of triples to include context information. But of course, there are again many properties with a lower number of languages, as the median is also only 6 languages. Cross-lingual EL is therefore certainly necessary to use language model-based EL in multiple languages.

By using the qualifiers of hyper-relational statements, more detailed information is available, useful not only for Entity Linking but also for other problems like Question Answering. The inclusion of hyper-relational statements is also more challenging. Novel graph embeddings have to be developed and utilized, which can represent the structure of a claim enriched with qualifiers [37,98].

[13] https://forum.dbpedia.org/t/differences-in-results-on-dbpedia-live-and-http-dbpedia-org-sparql-endpoint/888/2

[14] There exist some predicates (e.g., `foaf:name`, `dbp:commonName` or `dbp:conventionalLongName`) that might point to aliases but they are often either not used or specify already stated aliases.

Table 5
Number of English labels/aliases pointing to a certain number of items in Wikidata (extracted from dump [125])

# Labels/aliases	70,124,438	2,041,651	828,471	89,210	3329
# Items per label/alias	1	2	3–10	11–100	<100

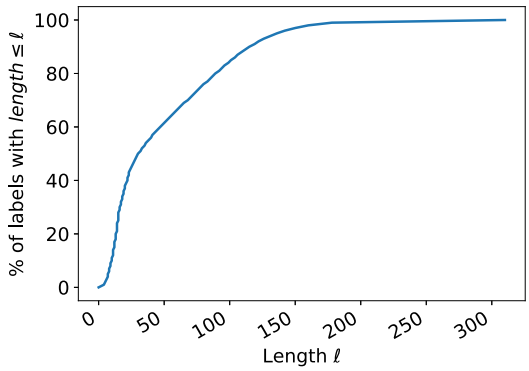


Fig. 7. Percentiles of English label lengths (extracted from dump [125]).

Ranks are of use for EL in the following way. Imagine a person had multiple spouses throughout his/her life. In Wikidata, all those relationships are assigned to the person via statements of different ranks. If now an utterance is encountered containing information on the person and her/his spouse, one can utilize the Wikidata statements for comparison. Depending on the time point of the utterance, different statements apply. One could, for example, weigh the relevance of statements according to their rank. If now a KG (for example Yago4 [109]) includes only the most valid statement, the current spouse, utterances containing past spouses are harder to link.

For references, up to now, no found approach did utilize them for EL. One use case might be to filter statements by reference if one knows the source's credibility, but this is more a measure to cope with the uncertainty of statements in Wikidata and not directly related to EL.

Weaknesses However, this community-driven approach also introduces challenges. For example, the list of labels of an item will not be exhaustive, as shown in Figs 6b and 6c. The graphs consider labels and aliases of all languages. While the median of labels and aliases is around 4 per element, not all are useful for Entity Linking. `Ennio Morricone` does not have an alias solely consisting of `Ennio` while he will certainly sometimes be referenced by that. Thus, one can not rely on the exact labels alone. But interestingly, Wikidata has properties for the fore- and surname alone, just not as a label or alias. A close examination of what information to use is essential.

This is also a problem in other KGs. Also, Wikidata often has items with very long, noisy, error-prone labels, which can be a challenge to link to [78]. Nearly 20 percent of labels have a length larger than 100 letters, see Fig. 7. Due to the community-driven approach, false statements also occur due to errors or vandalism [47].

Another problem is that entities lack of facts (here defined as statements not being labels, descriptions, or aliases). According to Tanon et al. [109], in March 2020, DBpedia had, on average, 26 facts per entity while Wikidata had only 12.5. This is still more than YAGO4 with 5.1. To tackle such long-tail entities, different approaches are necessary. The lack of descriptions can also be a problem. Currently, around 10% of all items do not have a description, as shown in Fig. 6d. Luckily, the situation is increasingly improving.

A general problem of Entity Linking is that a label or alias can reference multiple entities, see Table 5. While around 70 million mentions point each to a unique item, 2.9 million do not. Not all of those are entities by our definition but, e.g., also classes or topics. In addition, longer labels or aliases often correspond to non-entity items. Thus, the percentage of entities with overlapping labels or aliases is certainly larger than for all items. To use Wikidata as a Knowledge Graph, one needs to be cautious of the items one will include as entities. For example, there exist `Wikimedia disambiguation page` items that often have the same label as an entity in the classic sense. Both `Q76` and `Q61909968` have `Barack Obama` as the label. Including those will make disambiguation more difficult. Also, the possibility of contradictory facts will make EL over Wikidata harder.

In Wikification, also known as EL on Wikipedia, large text documents for each entity exist in the knowledge graph, enabling text-heavy methods [127]. Such large textual contexts (besides the descriptions and the labels of triples itself) do not exist in Wikidata, requiring other methods or the inclusion of Wikipedia. However, as Wikidata is closely related to Wikipedia, an inclusion is easily doable. Every Wikipedia article is connected to a Wikidata item. The Wikipedia article belonging to a Wikidata item can be, for example, extracted via a SPARQL[15] query to the Wikidata Query Service[16] using the http://schema.org/about predicate. The Wikidata item of a Wikipedia article can be simply found on the article page itself or by using the Wikipedia API.[17]

One can conclude that the characteristics of Wikidata, like being up to date, multilingual and hyper-relational, introduce new possibilities. At the same time, the existence of long-tail entities, noise or contradictory facts poses a challenge.

5. Datasets

5.1. Overview

This section is concerned with analyzing the different datasets which are used for Wikidata EL. A comparison can be found in Table 6. The majority of datasets on which existing Entity linkers were evaluated, were originally constructed for KGs different from Wikidata. Such a mapping can be problematic as some entities labeled for other KGs could be missing in Wikidata. Or some NIL entities that do not exist in other KGs could exist in Wikidata. Eleven datasets [16,23,24,27,29,33,46,56,69,80] were found for which Wikidata identifiers were available from the start. In the following the datasets are separated by their domain. A list of all examined datasets – including links where available – can be found in the Appendix in Table 17.

5.1.1. Encyclopedic datasets

LC-QuAD 2.0 [27] is a semi-automatically created dataset for Questions Answering providing complex natural language questions. For each question, Wikidata and DBpedia identifiers are provided. The questions are generated from subgraphs of the Wikidata KG and then manually checked. The dataset does not provide annotated mentions.

T-REx [33] was constructed automatically over Wikipedia abstracts. Its main purpose is Knowledge Base Population (KBP). According to Mulang et al. [78], this dataset describes the challenges of Wikidata, at least in the form of long, noisy labels, the best.

The Kensho Derived Wikimedia Dataset [56] is an automatically created condensed subset of Wikimedia data. It consists of three levels: Wikipedia text, annotations with Wikipedia pages and links to Wikidata items. Thus, mentions in Wikipedia articles are annotated with Wikidata items. However, as some Wikidata items do not have a corresponding Wikipedia page, the annotation is not exhaustive. It was constructed for NLP in general.

5.1.2. Research-focused datasets

ISTEX-1000 [24] is a research-focused dataset containing 1000 author affiliation strings. It was manually annotated to evaluate the OpenTapioca [24] entity linker.

5.1.3. Biographical datasets

KnowledgeNet [23] is a Knowledge Base Population dataset with 9073 manually annotated sentences. The text was extracted from biographical documents from the web or Wikipedia articles.

5.1.4. News datasets

NYT2018 [68,69] consists of 30 news documents that were manually annotated on Wikidata and DBpedia. It was constructed for KBPearl [69], so its main focus is also KBP which is a downstream task of EL.

One dataset, KORE50DYWC [80], was found, which was not used by any of the approach papers. It is an annotated EL dataset based on the KORE50 dataset, a manually annotated subset of the AIDA-CoNLL corpus. The

[15] https://www.w3.org/TR/rdf-sparql-query/
[16] https://query.wikidata.org
[17] https://en.wikipedia.org/w/api.php

Table 6
Comparison of used datasets

Dataset	Domain	Year	Annotation process	Purpose	Spans given	Identifiers
T-REx [33]	Wikipedia abstracts	2015	automatic	Knowledge Base Population (KBP), Relation Extraction (RE), Natural Language Generation (NLG)	✓	Wikidata
NYT2018 [68,69]	News	2018	manually	EL	✓	Wikidata, DBpedia
ISTEX-1000 [24]	Research articles	2019	manually	EL	✓	Wikidata
LC-QuAD 2.0 [27]	General complex questions (Wikidata)	2019	semi-automatic	Question Answering (QA)	×	DBpedia, Wikidata
Knowledge Net [23]	Wikipedia abstracts, biographical texts	2019	manually	KBP	✓	Wikidata
KORE50DYWC [80]	News	2019	manually	EL	✓	Wikidata, DBpedia, YAGO, Crunchbase
Kensho Derived Wikimedia Dataset [56]	Wikipedia	2020	automatic	Natural Language Processing (NLP)	✓	Wikidata, Wikipedia
CLEF HIPE 2020 [29]	Historical newspapers	2020	manually	ER, EL	✓	Wikidata
Mewsli-9 [16]	News in multiple languages	2020	automatic	Multilingual EL	✓	Wikidata
TweekiData [46]	Tweets	2020	automatic	EL	✓	Wikidata
TweekiGold [46]	Tweets	2020	manually	EL	✓	Wikidata

[1]Data from 2010
[2]Original dataset on Wikipedia

original KORE50 dataset focused on highly ambiguous sentences. All sentences were reannotated with DBpedia, Yago, Wikidata and Crunchbase entities.

CLEF HIPE 2020 [29] is a dataset based on historical newspapers in English, French and German. Only the English dataset will be analyzed in the following. This dataset is of great difficulty due to many errors in the text, which originate from the OCR method used to parse the scanned newspapers. For the English language, only a development and test set exist. In the other two languages, a training set is also available. It was manually annotated.

Mewsli-9 [16] is a multilingual dataset automatically constructed from WikiNews. It includes nine different languages. A high percentage of entity mentions in the dataset do not have corresponding English Wikipedia pages, and thus, cross-lingual linking is necessary. Again, only the English part is included during analysis.

5.1.5. Twitter datasets

TweekiData and TweekiGold [46] are an automatically annotated corpus and a manually annotated dataset for EL over tweets. TweekiData was created by using other existing tweet-based datasets and linking them to Wikidata data via the Tweeki EL. TweekiGold was created by an expert, manually annotating tweets from another dataset with Wikidata identifiers and Wikipedia page-titles.

5.2. Analysis

Table 7 shows the number of documents, the number of mentions, NIL entities and unique entities, and the mentioned ratio. What classifies as a document in a dataset depends on the dataset itself. For example, for T-REx, a document is a whole paragraph of a Wikipedia article, while for LC-QuAD 2.0, a document is just a single question. Due to this, the average number of entities in a document also varies, e.g., LC-QuAD 2.0 with 1.47 entities per document and T-REx with 11.03. If a dataset was not available, information from the original paper was included. If dataset splits were available, the statistics are also shown separately. The majority of datasets do not

Table 7
Comparison of the datasets with focus on the number of documents and Wikidata entities

Dataset	# documents	# mentions	NIL entities	Wikidata entities	Unique Wikidata entities	Mentions per document
T-REx [33]	4,650,000	51,297,484	0%	100%	9.1%	11.03
NYT2018 [68,69][1]	30	–	–	–	–	–
ISTEX-1000 [24] (train)	750	2073	0%	100%	53.7%	2.76
ISTEX-1000 [24] (test)	250	670	0%	100%	65.8%	2.68
LC-QuAD 2.0 [27]	6046	44,529	0%	100%	51.2%	1.47
Knowledge Net [23] (train)	3977	13,039	0%	100%	30%	3.28
Knowledge Net [23] (test)[2]	1014	–	–	–	–	–
KORE50DYWC [80]	50	307	0%	100%	72.0%	6.14
Kensho Derived Wikimedia Dataset [56]	14,255,258	121,835,453	0%	100%	3.7%	8.55
CLEF HIPE 2020 (en, dev) [29]	80	470	46.4%	53.6%	31.9%	5.88
CLEF HIPE 2020 (en, test) [29]	46	134	33.6%	66.4%	42.5%	2.91
Mewsli-9 (en) [16]	12,679	80,242	0%	100%	48.2%	6.33
TweekiData [46]	5,000,000	5,038,870	61.2%	38.8%	5.4%	1.01
TweekiGold [46]	500	958	11.1%	88.9%	66.6%	1.92

[1] Information gathered from accompanying paper as dataset was not available
[2] Available dataset did not contain mention/entity information

contain NIL entities. For the Tweeki datasets, it is not mentioned which Wikidata dump was used to annotate. For a dataset that contains NIL entities, this is problematic. On the other hand, the dump is specified for the CLEF HIPE 2020 dataset, making it possible to work on the Wikidata version with the correct entities missing.

Answer – RQ 1. ***Which Wikidata EL datasets exist, how widely used are they and how are they constructed?***

The preceding paragraphs answer the following two aspects of the first research question. First, we provided descriptions and an overview of all datasets created for Wikidata, including statistics on their structure. This answers which datasets exist. Furthermore, for each dataset it is stated how they were constructed, whether automatically, semi-automatically or manually. Thus information on the quality and construction process of the datasets is given. To answer the last part of the question, how widely are the datasets in use, Table 8 shows how many times each Wikidata dataset was used in Wikidata EL approaches during training or evaluation. As one can see, there exists no single dataset used in all research of EL. This is understandable as different datasets focus on different document types and domains as shown in Table 6, what again results in different approaches.

The difficulty of the different datasets was measured by the accuracy of a simple EL method (Table 10) and the ambiguity of mentions (Table 9). The simple EL method searches for entity candidates via an ElasticSearch index, including all English labels and aliases. It then disambiguates by taking the one with the largest tf-idf-based BM25 similarity measure score and the lowest Q-identifier number resembling the popularity. Nothing was done to handle inflections.[18] Only accessible datasets were included. As one can see, the accuracy is positively correlated with the number of exact matches. The more ambiguous the underlying entity mentions are, the more inaccurate a simple similarity measure between label and mention becomes. In this case, more context information is necessary. The simple Entity Linker was only applied to datasets that were feasible to disambiguate in that way. T-REx and the Kensho Derived Wikimedia Dataset were too large in terms of the number of documents to run the linker on commodity hardware. According to the EL performance, ISTEX-1000 is the easiest dataset. Many of the ambiguous mentions reference the most popular one, while also many exact unique matches exist. T-REx, the Kensho Derived

[18] All source code, plots and results can be found on https://github.com/semantic-systems/ELEnglishWD.

Table 8
Usage of datasets for training or evaluation

Dataset	Number of usages in Wikidata EL approach papers
T-REx [33]	2 [69,78]
NYT2018 [68,69]	1 [69]
ISTEX-1000 [24]	2 [24,77]
LC-QuAD 2.0 [27]	2 [8,100]
Knowledge Net [23]	1 [69]
KORE50DYWC [80]	0
Kensho Derived Wikimedia Dataset [56]	1 [86]
CLEF HIPE 2020 [29]	3 [15,65,89]
Mewsli-9 [16]	1 [16]
TweekiData [46]	1 [46]
TweekiGold [46]	1 [46]

Table 9
Ambiguity of mentions (existence of a match does not correspond to a correct match), NYT2018 dataset was not available and LC-QuAD 2.0 is not annotated

Dataset	Average number of matches	No match	Exact match	More than one match
T-REx	4.79	31.36%	32.98%	35.65%
ISTEX-1000 (train)	23.23	8.06%	26.34%	65.61%
ISTEX-1000 (test)	25.85	10.30%	23.88%	65.82%
Knowledge Net (train)	21.90	10.41%	22.29%	67.3%
KORE50DYWC	28.31	3.93%	7.49%	88.60%
Kensho Derived Wikimedia Dataset	8.16	35.18%	30.94%	33.88%
CLEF HIPE 2020 (en, dev)	24.02	35.71%	11.51%	52.78%
CLEF HIPE 2020 (en, test)	17.78	43.82%	6.74%	49.44%
Mewsli-9 (en)	11.09	16.80%	34.90%	47.30%
TweekiData	19.61	19.98%	12.01%	68.01%
TweekiGold	16.02	7.41%	20.25%	72.34%

Wikimedia Dataset and the Mewsli-9 training dataset have the largest percentage of exact matches for labels. While TweekiGold is quite ambiguous, deciding on the most prominent entity appears to produce good EL results. The most ambiguous dataset is KORE50DYWC. Additionally, just choosing the most popular entity of the exact matches results in worse performs than for example on TweekiGold which is also very ambiguous. This is due to the fact that the original KORE50 dataset focuses on difficult ambiguous entities which are not necessarily popular. The CLEF HIPE 2020 dataset also has a low EL accuracy but not due to ambiguity but many mentions with no exact match. The reason for that is the noise created by OCR.

The second column of Table 10 specifies the accuracy with all unique exact matches removed. This is based on the intuition that exact matches without any competitors are usually correct.

As seen in the Tables 6, 7, 9 and 10, there exists a very diverse set of datasets for EL on Wikidata, differing in the domain, document type, ambiguity and difficulty.

Answer – RQ 2. ***Do the characteristics of Wikidata matter for the design of EL datasets and if so, how?***

Except the Mewsli-9 [16] and CLEF HIPE 2020 [29] datasets, none of the others take any specific characteristics of Wikidata into account. The two exceptions focus on multilinguality and rely therefore directly on the language-agnostic nature of Wikidata. The CLEF HIPE 2020 dataset is designed for Wikidata and has documents for English, French and German, but each language has a different corpus of documents. The same is the case for the Mewsli-9 dataset, while here, documents in nine languages are available. In the future, a dataset similar to VoxEL [96], which is defined for Wikipedia, would be helpful. Here, each utterance is translated into multiple languages, which eases the comparison of the multilingual EL performance. Having the same corpus of documents in different languages

Table 10
EL accuracy – Kensho derived Wikimedia dataset, T-REx and TweekiData are not included due to size, **Acc. filtered** has all exact matches removed, NYT2018 dataset was not available and LC-QuAD 2.0 is not annotated

Dataset	Acc.	Acc. filtered
ISTEX-1000 (train)	0.744	0.716
ISTEX-1000 (test)	0.716	0.678
Knowledge Net (train)	0.371	0.285
KORE50DYWC	0.225	0.187
CLEF HIPE 2020 (en, dev)	0.333	0.287
CLEF HIPE 2020 (en, test)	0.258	0.241
TweekiGold	0.565	0.520
Mewsli-9 (en)	0.602	0.490

would allow a better comparison of a method's performance in various languages. Of course, such translations will never be perfectly comparable.

Besides that, we identified one additional characteristic which might be of relevance to Wikidata EL datasets. It is the large rate of change of Wikidata. Due to that, it would be advisable that the datasets specify the Wikidata dumps they were created on, similar to Petroni et al. [88]. Many of the existing datasets do that, yet not all. In current dumps, entities, which were available while the dataset was created, could have been removed. It is even more probable that NIL entities could now have a corresponding entity in an updated Wikidata dump version. If the EL approach now would detect it as a NIL entity, it is evaluated as correct, but in reality, this is false and vice versa. Of course, this is not a problem unique to Wikidata. Anytime, the dump is not given for an EL dataset, similar uncertainties will occur. But due to the fast growth of Wikidata (see Fig. 6a), this problem is more pronounced.

Concerning *emerging entities*, another variant of an EL dataset could be useful too. Two Wikidata dumps from different time points could be used to label the utterances. Such a dataset would be valuable in the context of an EL approach supporting emerging entities (e.g., the approach by Hoffart et al. [50]). With the true entities available, one could measure the quality of the created emerging entities. That is, multiple mentions assigned to the same emerging entity should also point to a single entity in the more recent KG. Also, constraining that the method needs to perform well on both KG dumps would force EL approaches to be less reliant on a fixed graph structure.

6. Approaches

Currently, the number of methods intended to work explicitly on Wikidata is still relatively small, while the amount of the ones utilizing the characteristics of Wikidata is even smaller.

There exist several KG-agnostic EL approaches [76,114,137]. However, they were omitted as their focus is being independent of the KG. While they are able to use Wikidata characteristics like labels or descriptions, there is no explicit usage of those. They are available in most other KGs. None of the found KG-agnostic EL papers even mentioned Wikidata. Though we recognize that KG-agnostic approaches are very useful in the case that a KG becomes obsolete and has to be replaced or a non-public KG needs to be used, such approaches are not included in this section. However, Table 15 in the Appendix provides an overview of the used Wikidata characteristics of the three approaches.

DeepType [90] is an entity linking approach relying on the fine-grained type system of Wikidata and the categories of Wikipedia. As type information is not evolving as fast as novel entities appear, it is relatively robust against a changing knowledge base. While it uses Wikidata, it is not specified in the paper whether it links to Wikipedia or Wikidata. Even the examination of the available code did not result in an answer as it seems that the entity linking component is missing. While DeepType showed that the inclusion of Wikidata type information is very beneficial in entity linking, we did not include it in this survey due to the aforementioned reasons. As Wikidata contains many more types (≈2,400,000) than other KGs, e.g., DBpedia (≈484,000) [109][19]), it seems to be more suitable for this

[19] If all rdf:type objects are considered, else ≈ 768 (gathered via https://dbpedia.org/sparql/) if only considering types of the DBpedia ontology.

fine-grained type classification. Yet, not only the number of types plays a role but also how many types are assigned per entity. In this regard, Wikipedia provides much more type information per entity than Wikidata [124]. That is probably the reason why both Wikipedia categories and Wikidata types are used together. As Wikidata is growing every minute, it may also be challenging to keep the type system up to date.

Tools without accompanying publications are not considered due to the lack of information about the approach and its performance. Hence, for instance, the Entity Linker in the DeepPavlov [17] framework is not included, although it targets Wikidata and appears to use label and description information successfully to link entities.

While the approach by Zhou et al. [136] does utilize Wikidata aliases in the candidate generation process, the target KB is Wikipedia and was therefore excluded.

The vast majority of methods is using machine learning to solve the EL task [8,15,16,18,24,53,60,65,77,78,86,89,105]. Some of those approaches solve the ER and EL jointly as an end-to-end task. Besides that, there exist two rule-based approaches [46,100] and two based on graph optimization [60,69].

The approaches mentioned above solve the EL problem as specified in Section 3. That is, other EL methods with a different problem definition also exist. For example, Almeida et al. [4] try to link street names to entities in Wikidata by using additional location information and limiting the entities only to locations. As it uses additional information about the true entity via the location, it is less comparable to the other approaches and, thus, was excluded from this survey. Thawani et al. [111] link entities only over columns of tables. The approach is not comparable since it does not use natural language utterances. The approach by Klie et al. [62] is concerned with Human-In-The-Loop EL. While its target KB is Wikidata, the focus on the inclusion of a human in EL process makes it incomparable to the other approaches. EL methods exclusively working on languages other than English [30–32,59,116] were not considered but also did not use any novel characteristics of Wikidata. In connection to the CLEF HIPE 2020 challenge [30], multiple Entity Linkers working on Wikidata were built. While short descriptions of the approaches are available in the challenge-accompanying paper, only approaches described in an own published paper were included in this survey. The approach by Kristanti and Romary [64] was not included as it used pre-existing tools for EL over Wikidata, for which no sufficient documentation was available.

Due to the limited number of methods, we also evaluated methods that are not solely using Wikidata but also additional information from a separate KG or Wikipedia. This is mentioned accordingly. Approaches linking to knowledge graphs different from Wikidata, but for which a mapping between the knowledge graphs and Wikidata exists, are also not included. Such methods would not use the Wikidata characteristics at all, and their performance depends on the quality of the other KG and the mapping.

In the following, the different approaches are described and examined according to the used characteristics of Wikidata. An overview can be found in Table 11. We split the approaches into two categories, the ones doing only EL and the ones doing ER and EL. Furthermore, to provide a better overview of the existing approaches, they are categorized by notable differences in their architecture or used features. This categorization focuses on the EL aspect of the approaches.

For each approach, it is mentioned what datasets were used in the corresponding paper. Only a subset of the datasets was directly annotated with Wikidata identifiers. Hence, datasets are mentioned, which do not occur in Section 5.

6.1. Entity linking

6.1.1. Language model-based approaches

The approach by Mulang et al. [77] is tackling the EL problem with transformer models [117]. It is assumed that the candidate entities are given. For each entity, the labels of 1-hop and 2-hop triples are extracted. Those are then concatenated together with the utterance and the entity mention. The concatenation is the input of a pre-trained transformer model. With a fully connected layer on top, it is then optimized according to a binary cross-entropy loss. This architecture results in a similarity measure between the entity and the entity mention. The examined models are the transformer models Roberta [72], XLNet [131] and the DCA-SL model [130]. The approach was evaluated on three datasets with no focus on certain documents or domains: ISTEX-1000 [24], Wikidata-Disamb [18] and AIDA-CoNLL [51]. AIDA-CoNLL is a popular dataset for evaluating EL but has Wikipedia as the target. ISTEX-1000 focuses on research documents, and Wikidata-Disamb is an open-domain dataset. There is no global coherence

Table 11
Comparison between the utilized Wikidata characteristics of each approach

Approach	Labels/Aliases	Descriptions	Knowledge graph structure	Hyper-relational structure	Types	Additional information
OpenTapioca [24]	✓	×	✓	✓	✓	×
Falcon 2.0 [100]	✓	×	✓[1]	×	×	×
Arjun [78]	✓	×	×	×	×	×
VCG [105]	✓	×	✓	×	×	×
KBPearl [69]	✓	×	✓	×	×	×
PNEL [8]	✓	✓	✓	×	×	×
Mulang et al. [77]	✓	✓[2]	✓	×	×	×
Perkins [86]	✓	×	✓	×	×	×
NED using DL on Graphs [18]	✓	×	✓	×	×	×
Huang et al. [53]	✓	✓	✓	×	×	Wikipedia
Boros et al. [15]	×	×	×	×	✓	Wikipedia, DBpedia
Provatorov et al. [89]	✓	✓	×	×	×	Wikipedia
Labusch and Neudecker [65]	×	×	×	×	×	Wikipedia
Botha et al. [16]	×	×	×	×	×	Wikipedia
Hedwig [60]	✓	✓	✓	×	×	Wikipedia
Tweeki [46]	✓	×	×	×	✓	Wikipedia

[2] Appears in the set of triples used for disambiguation
[1] Only querying the existence of triples

technique applied. Overall, up to 2-hop triples of any kind are used. For example, labels, aliases, descriptions, or general relations to other entities are all incorporated. It is not mentioned if the hyper-relational structure in the form of qualifiers was used. On the one hand, the purely language-based EL results in less need for retraining if the KG changes as shown by other approaches [16,127]. This is the case due to the reliance on sub-word embeddings and pre-training via the chosen transformer models. If full word-embeddings were used, the inclusion of new words would make retraining necessary. Still, an evaluation of the model on the zero-shot EL task is missing and has to be done in the future. The reliance on the triple information might be problematic for long-tail entities which are rarely referred to and are part of fewer triples. Nevertheless, a lack of available context information is challenging for any EL approach relying on it.

The approach designed by Botha et al. [16] tackles multilingual EL. It is also crosslingual. That means it can link entity mentions to entities in a knowledge graph in a language different from the utterance one. The idea is to train one model to link entities in utterances of 100+ different languages to a KG containing not necessarily textual information in the language of the utterance. While the target KG is Wikidata, they mainly use Wikipedia descriptions as input. This is the case as extensive textual information is not available in Wikidata. The approach resembles the Wikification method by Wu et al. [127] but extends the training process to be multilingual and targets Wikidata. Candidate generation is done via a dual-encoder architecture. Here, two BERT-based transformer models [26] encode both the context-sensitive mentions and the entities to the same vector space. The mentions are encoded using local context, the mention and surrounding words, and global context, the document title. Entities are encoded by using the Wikipedia article description available in different languages. In both cases, the encoded CLS-token are projected to the desired encoding dimension. The goal is to embed mentions and entities in such a way that the embeddings are similar. The model is trained over Wikipedia by using the anchors in the text as entity mentions. There exists no limitation that the used Wikipedia articles have to be available in all supported languages. If an article is missing in the English Wikipedia but available in the German one, it is still included. Now, after the model is trained, all entities are embedded. The candidates are generated by embedding the mention and searching for the nearest neighbors. A cross-encoder is employed to rank the entity candidates, which cross-encodes entity description and mention text together by concatenating and feeding them into a BERT model. Final scores are obtained, and the entity mention is linked. The model was evaluated on the cross-lingual EL dataset TR2016^{hard} [112] and the multilingual EL dataset Mewsli-9 [16]. Furthermore, it was tested how well it performs on an English-only

dataset called WikiNews-2018 [42]. Wikidata information is only used to gather all the Wikipedia descriptions in the different languages for all entities. The approach was tested on zero- and few-shot settings showing that the model can handle an evolving knowledge graph with newly added entities that were never seen before. This is also more easily achievable due to its missing reliance on the graph structure of Wikidata or the structure of Wikipedia. It is the case that some Wikidata entities do not appear in Wikipedia and are therefore invisible to the approach. But as the model is trained on descriptions of entities in multiple languages, it has access to many more entities than only the ones available in the English Wikipedia.

6.1.2. Language model and graph embeddings-based approaches

The master thesis by Perkins [86] is performing candidate generation by using anchor link probability over Wikipedia and locality-sensitive hashing (LSH) [43] over labels and mention bi-grams. Contextual word embeddings of the utterance (ELMo [87]) are used together with KG embeddings (TransE [14]), calculated over Wikipedia and Wikidata, respectively. The context embeddings are sent through a recurrent neural network. The output is concatenated with the KG embedding and then fed into a feed-forward neural network resulting in a similarity measure between the KG embedding of the entity candidate and the utterance. It was evaluated on the AIDA-CoNLL [51] dataset. Wikidata is used in the form of the calculated TransE embeddings. Hyper-relational structures like qualifiers are not mentioned in the thesis and are not considered by the TransE embedding algorithm and, thus, probably not included. The used KG embeddings make it necessary to retrain when the Wikidata KG changes as they are not dynamic.

6.1.3. Word and graph embeddings-based approaches

In 2018, Cetoli et al. [18] evaluated how different types of basic neural networks perform solely over Wikidata. Notably, they compared the different ways to encode the graph context via neural methods, especially the usefulness of including topological information via GNNs [106,129] and RNNs [49]. There is no candidate generation as it was assumed that the candidates are available. The process consists of combining text and graph embeddings. The text embedding is calculated by applying a Bi-LSTM over the Glove Embeddings of all words in an utterance. The resulting hidden states are then masked by the position of the entity mention in the text and averaged. A graph embedding is calculated in parallel via different methods utilizing GNNs or RNNs. The end score is the output of one feed-forward layer having the concatenation of the graph and text embedding as its input. It represents if the graph embedding is consistent with the text embedding. Wikidata-Disamb30 [18] was used for evaluating the approach. Each example in the dataset also contains an ambiguous negative entity, which is used during training to be robust against ambiguity. One crucial problem is that those methods only work for a single entity in the text. Thus, it has to be applied multiple times, and there will be no information exchange between the entities. While the examined algorithms do utilize the underlying graph of Wikidata, the hyper-relational structure is not taken into account. The paper is more concerned with comparing how basic neural networks work on the triples of Wikidata. Due to the pure analytical nature of the paper, the usefulness of the designed approaches to a real-world setting is limited. The reliance on graph embeddings makes it susceptible to change in the Wikidata KG.

6.2. Entity recognition and entity linking

The following methods all include ER in their EL process.

6.2.1. Language model-based approaches

In connection to the *CLEF 2020 HIPE challenge* [30], multiple approaches [15,65,89] for ER and EL of historical newspapers on Wikidata were developed. Documents were available in English, French and German. Three approaches with a focus on the English language are described in the following. Differences in the usage of Wikidata between the languages did not exist. Yet, the approaches were not multilingual as different models were used and/or retraining was necessary for different languages.

Boros et al. [15] tackled ER by using a BERT model with a CRF layer on top, which recognizes the entity mentions and classifies the type. During the training, the regular sentences are enriched with misspelled words to make the model robust against noise. For EL, a knowledge graph is built from Wikipedia, containing Wikipedia titles, page ids, disambiguation pages, redirects and link probabilities between mentions and Wikipedia pages are calculated. The link probability between anchors and Wikipedia pages is used to gather entity candidates for a mention.

The disambiguation approach follows an already existing method [63]. Here, the utterance tokens are embedded via a Bi-LSTM. The token embeddings of a single mention are combined. Then similarity scores between the resulting mention embedding and the entity embeddings of the candidates are calculated. The entity embeddings are computed according to Ganea and Hofmann [39]. These similarity scores are combined with the link probability and long-range context attention, calculated by taking the inner product between an additional context-sensitive mention embedding and an entity candidate embedding. The resulting score is a local ranking measure and is again combined with a global ranking measure considering all other entity mentions in the text. In the end, additional filtering is applied by comparing the DBpedia types of the entities to the ones classified during the ER. If the type does not match or other inconsistencies apply, the entity candidate gets a lower rank. Here, they also experimented with Wikidata types, but this resulted in a performance decrease. As can be seen, technically, no Wikidata information besides the unsuccessful type inclusion is used. Thus, the approach resembles more of a Wikification algorithm. Yet, they do link to Wikidata as the HIPE task dictates it, and therefore, the approach was included in the survey. New Wikipedia entity embeddings can be easily added [39] which is an advantage when Wikipedia changes. Also, its robustness against erroneous texts makes it ideal for real-world use. This approach reached SOTA performance on the CLEF 2020 HIPE challenge.

Labusch and Neudecker [65] also applied a BERT model for ER. For EL, they used mostly Wikipedia, similar to Boros et al. [15]. They built a knowledge graph containing all person, location and organization entities from the German Wikipedia. Then it was converted to an English knowledge graph by mapping from the German Wikipedia Pages via Wikidata to the English ones. This mapping process resulted in the loss of numerous entities. The candidate generation is done by embedding all Wikipedia page titles in an Approximative Nearest Neighbour index via BERT. Using this index, the neighboring entities to the mention embedding are found and used as candidates. For ranking, anchor-contexts of Wikipedia pages are embedded and fed into a classifier together with the embedded mention-context, which outputs whether both belong to the same entity. This is done for each candidate for around 50 different anchor contexts. Then, multiple statistics on those similarity scores and candidates are calculated, which are used in a Random Forest model to compute the final ranks. Similar to the previous approach, Wikidata was only used as the target knowledge graph, while information from Wikipedia was used for all the EL work. Thus, no special characteristics of Wikidata were used. The approach is less affected by a change of Wikidata due to similar reasons as the previous approach. This approach lacks performance compared to the state of the art in the HIPE task. The knowledge graph creation process produces a disadvantageous loss of entities, but this might be easily changed.

Provatorov et al. [89] used an ensemble of fine-tuned BERT models for ER. The ensemble is used to compensate for the noise of the OCR procedure. The candidates were generated by using an ElasticSearch index filled with Wikidata labels. The candidate's final rank is calculated by taking the search score, increasing it if a perfect match applies and finally taking the candidate with the lowest Wikidata identifier number (indicating a high popularity score). They also created three other methods of the EL approach: (1) The ranking was done by calculating cosine similarity between the embedding of the utterance and the embedding of the same utterance with the mention replaced by the Wikidata description. Furthermore, the score is increased by the Levenshtein distance between the entity label and the mention. (2) A variant was used where the candidate generation is enriched with historical spellings of Wikidata entities. (3) The last variant used an existing tool [115], which included contextual similarity and co-occurrence probabilities of mentions and Wikipedia articles. In the tool, the final disambiguation is based on the ment-norm method by Le and Titov [66]. The approach uses Wikidata labels and descriptions in one variant of candidate ranking. Beyond that, no other characteristics specific to Wikidata were considered. Overall, the approach is very basic and uses mostly pre-existing tools to solve the task. The approach is not susceptible to a change of Wikidata as it is mainly based on language and does not need retraining.

The approach designed by Huang et al. [53] is specialized in short texts, mainly questions. The ER is performed via a pre-trained BERT model [26] with a single classification layer on top, determining if a token belongs to an entity mention. The candidate search is done via an ElasticSearch[20] index, comparing the entity mention to labels and aliases by exact match and Levenshtein distance. The candidate ranking uses three similarity measures

[20] https://www.elastic.co/elasticsearch/

to calculate the final rank. A CNN is used to compute a character-based similarity between entity mention and candidate label. This results in a similarity matrix whose entries are calculated by the cosine similarity between each character embedding of both strings. The context is included in two ways. First, between the utterance and the entity description, by embedding the tokens of each sequence through a BERT model. Again, a similarity matrix is built by calculating the cosine similarity between each token embedding of both utterance and description. The KG is also considered by including the triples containing the candidate as a subject. For each such triple, a similarity matrix is calculated between the label concatenation of the triple and the utterance. The most representative features are then extracted out of the matrices via max-pooling, concatenated and fed into a two-layer perceptron. The approach was evaluated on the WebQSP [105] dataset, which is composed of short questions from web search logs. Wikidata labels, aliases and descriptions are utilized. Additionally, the KG structure is incorporated through the labels of candidate-related triples. This is similar to the approach by Mulang et al. [77], but only 1-hop triples are used. There is also no hyper-relational information considered. Due to its reliance on text alone and using a pre-trained language model with sub-word embeddings, it is less susceptible to changes of Wikidata. While the approach was not empirically evaluated on the zero-shot EL task, other approaches using language models (LM) [16,73,127] were and indicate a good performance.

6.2.2. Word embedding-based approaches

Arjun [78] tries to tackle specific challenges of Wikidata like long entity labels and implicit entities. Published in 2020, Arjun is an end-to-end approach utilizing the same model for ER and EL. It is based on an Encoder-Decoder-Attention model. First, the entities are detected via feeding Glove [85] embedded tokens of the utterance into the model and classifying each token as being an entity or not. Afterward, candidates are generated in the same way as in Falcon 2.0 [100] (see Section 6.2.6). The candidates are then ranked by feeding the mention, the entity label, and its aliases into the model and calculating the score. The model resembles a similarity measure between the mention and the entity labels. Arjun was trained and evaluated on the T-REx [33] dataset consisting of extracts out of various Wikipedia articles. It does not use any global ranking. Wikidata information is used in the form of labels and aliases in the candidate generation and candidate ranking. The model was trained and evaluated using GloVe embeddings, for which new words are not easily addable. New entities are therefore not easily supported. However, the authors claim that one can replace them with other embeddings like BERT-based ones. While those proved to perform quite well in zero-shot EL [16,127], this was usually done with more context information besides labels. Therefore it remains questionable if using those would adapt the approach for zero-shot EL.

6.2.3. Word and graph embeddings-based approaches

In 2018, Sorokin and Gurevych [105] were doing joint end-to-end ER and EL on short texts. The algorithm tries to incorporate multiple context embeddings into a mention score, signaling if a word is a mention, and a ranking score, signaling the candidate's correctness. First, it generates several different tokenizations of the same utterance. For each token, a search is conducted over all labels in the KG to gather candidate entities. If the token is a substring of a label, the entity is added. Each token sequence gets then a score assigned. The scoring is tackled from two sides. On the utterance side, a token-level context embedding and a character-level context embedding (based on the mention) are computed. The calculation is handled via dilated convolutional networks (DCNN) [133]. On the KG side, one includes the labels of the candidate entity, the labels of relations connected to a candidate entity, the embedding of the candidate entity itself, and embeddings of the entities and relations related to the candidate entity. This is again done by DCNNs and, additionally, by fully connected layers. The best solution is then found by calculating a ranking and mention score for each token for each possible tokenization of the utterance. All those scores are then summed up into a global score. The global assignment with the highest score is then used to select the entity mentions and entity candidates. The question-based EL datasets WebQSP [105] and GraphQuestions [108] were used for evaluation. GraphQuestions contains multiple paraphrases of the same questions and is used to test the performance on different wordings. The approach uses the underlying graph, label and alias information of Wikidata. Graph information is used via connected entities and relations. They also use TransE embeddings, and therefore no hyper-relational structure. Due to the usage of static graph embeddings, retraining will be necessary if Wikidata changes.

PNEL [8] is an end-to-end (E2E) model jointly solving ER and EL focused on short texts. PNEL employs a Pointer network [118] working on a set of different features. An utterance is tokenized into multiple different combinations. Each token is extended into the (1) token itself, (2) the token and the predecessor, (3) the token and the successor, and (4) the token with both predecessor and successor. For each token combination, candidates are searched for by using the BM25 similarity measure. Fifty candidates are used per tokenization combination. Therefore, 200 candidates (not necessarily 200 distinct candidates) are found per token. For each candidate, features are extracted. Those range from the simple length of a token to the graph embeddings of the candidate entity. All features are concatenated to a large feature vector. Therefore, per token, a sequence of 200 such features vectors exists. Finally, the concatenation of those sequences of each token in the sentence is then fed into a Pointer network. At each iteration of the Pointer network, it points to one distinct candidate in the network or an `END` token marking no choice. Pointing is done by computing a softmax distribution and choosing the candidate with the highest probability. Note that the model points to a distinct candidate, but this distinct candidate can occur multiple times. Thus, the model does not necessarily point to only a single candidate of the 200 ones. PNEL was evaluated on several QA datasets, namely WebQSP [105], SimpleQuestions [13] and LC-QuAD 2.0 [27]. SimpleQuestions focuses, as the name implies, on simple questions containing only very few entities. LC-QuAD 2.0, on the other hand, contains both, simple and more complex, longer questions including multiple entities. The entity descriptions, labels and aliases are all used. Additionally, the graph structure is included by TransE graph embeddings, but no hyper-relational information was incorporated. E2E models can often improve the performance of the ER. Most EL algorithms employed in the industry often use older ER methods decoupled from the EL process. Thus, such an E2E EL approach can be of use. Nevertheless, due to its reliance on static graph embeddings, complete retraining will be necessary if Wikidata changes.

6.2.4. Non-NN ML-based approaches

OpenTapioca [24] is a mainly statistical EL approach published in 2019. While the paper never mentions ER, the approach was evaluated with it. In the code, one can see that the ER is done by a SolrTextTagger analyzer of the Solr search platform.[21] The candidates are generated by looking up if the mention corresponds to an entity label or alias in Wikidata stored in a Solr collection. Entities are filtered out which do not correspond to the type person, location or organization. OpenTapioca is based on two main features, which are local compatibility and semantic similarity. First, local compatibility is calculated via a popularity measure and a unigram similarity measure between entity label and mention. The popularity measure is based on the number of sitelinks, PageRank scores, and the number of statements. Second, the semantic similarity strives to include context information in the decision process. All entity candidates are included in a graph and are connected via weighted edges. Those weights are calculated via a statistical similarity measure. This measure includes how likely it is to jump from one entity candidate to another while discounting it by the distance between the corresponding mentions in the utterance. The resulting adjacency matrix is then normalized to a stochastic matrix that defines a Markov Chain. One now propagates the local compatibility using this Markov Chain. Several iterations are then taken, and a final score is inferred via a Support Vector Machine. It supports multiple entities per utterance. OpenTapioca is evaluated on AIDA-CoNLL [51], Microposts 2016 [121], ISTEX-1000 [24] and RSS-500 [94]. RSS-500 consists of news-based examples and Microposts 2016 focuses on shorter documents like tweets. OpenTapioca was therefore evaluated on many different types of documents. The approach is only trained on and evaluated for three types of entities: locations, persons, and organizations. It facilitates Wikidata-specific labels, aliases, and sitelinks information. More importantly, it also uses qualifiers of statements in the calculation of the PageRank scores. But the qualifiers are only seen as additional edges to the entity. The usage in special domains is limited due to its restriction to only three types of entities, but this is just an artificial restriction. It is easily updatable if the Wikidata graph changes as no immediate retraining is necessary.

6.2.5. Graph optimization-based approaches

Hedwig [60] is a multilingual entity linker specialized on the TAC 2017 [55] task but published in 2020. Another entity linker [58], developed by the same authors, is not included in this survey as Hedwig is partly an evolution of it. The entities to be linked are limited to only a subset of all possible entity classes. Hedwig employs Wikidata and

[21] https://lucene.apache.org/solr/

Wikipedia at the same time. The Entity Recognition uses word2vec embeddings [75], character embeddings, and dictionary features where the character embeddings are calculated via a Bi-LSTM. The dictionary features are class-dependent, but this is not defined in more detail. Those embeddings and features are computed and concatenated for each token. Afterward, the whole sequence of token features is fed into a Bi-LSTM with a linear chain Conditional Random Field (CRF) layer at the end to recognize the entities. The candidates for each detected entity mention are then generated by using a mention dictionary. The dictionary is created from Wikidata and Wikipedia information, utilizing labels, aliases, titles or anchor texts. The candidates are disambiguated by constructing a graph consisting of all candidate entities, mentions, and occurring words in the utterance. The edges between entities and other entities, words, or mentions have the normalized pointwise mutual information (NPMI) assigned as their weights. The NPMI specifies how frequently two entities, an entity and a mention or an entity and a word, occur together. Those scores are calculated over a Wikipedia dump. Finally, the PageRank of each node in the graph is calculated via power iteration, and the highest-scoring candidates are chosen. The type classification is used to determine the types of entities, not mentions. As this is only relevant for the TAC 2017 task, the classifier can be ignored. The approach was evaluated on the TAC 2017 [55] dataset, which focuses on entities of type person, organization, location, geopolitics and facilities. The documents originate from discussion forums and newswire texts. Labels and aliases from multiple languages are used. It also uses sitelinks to connect the Wikidata identifiers and Wikipedia articles. The paper also claims to use descriptions but does not describe anywhere in what way. No hyper-relational or graph features are used. As it employs class-dependent features, it is limited to the entities of classes specified in the TAC 2017 task. The NPMI weights have to be updated with the addition of new elements in Wikidata and Wikipedia.

KBPearl [69], published in 2020, utilizes EL to populate incomplete KGs using documents. First, a document is preprocessed via Tokenization, POS tagging, NER, noun-phrase chunking, and time tagging. Also, an existing Information Extraction tool is used to extract open triples from the document. They experimented with four different tools (ReVerb [34], MinIE [41], ClausIE [21] and Stanford Open IE Tool [5]), Open triples are non-linked triples extracted via an open information extraction tool. The triples consist of a subject, predicate and object in unstructured text. For example, the open triple `<Ennio Morricone, composed, soundtrack of The Hateful Eight>` can be extracted from "Ennio Morricone, known for numerous famous soundtracks of the Spaghetti Western era, composed the soundtrack of the movie The Hateful Eight.". The triples are processed further by filtering invalid tokens and doing canonicalization. Then, a graph of entities, predicates, noun phrases, and relation phrases is constructed. The candidates are generated by comparing the noun/relation phrases to the labels and aliases of the entities/predicates. The edges between the entities/relations and between entities and relations are weighted by the number of intersecting one-hop statements. The next step is the computation of a maximum dense subgraph. Density is defined by the minimum weighted degree of all nodes [51]. As this problem is NP-hard, a greedy algorithm is used for optimization. New entities relevant for the task of Knowledge Graph Population are identified by thresholding the weighted sum of an entity's incident edges. Like used here, global coherence can perform sub-optimally since not all entities/relations in a document are related. Thus, two variants of the algorithm are proposed. First, a pipeline version that separates the full document into sentences. Second, a near neighbor mode, limiting the interaction of the nodes in the graph by the distances of the corresponding noun-phrases and relation-phrases. KBPearl was evaluated on many different datasets: ReVerb38 [69], NYT2018 [68,69], LC-QuAD 2.0 [27], QALD-7-WIKI [113], T-REx [33], Knowledge Net [23] and CC-DBP [44]. These datasets encompass news articles, questions, and general open-domain documents. The approach includes label and alias information of entities and predicates. Additionally, one-hop statement information is used, but hyper-relational features are not mentioned. However, the paper does not claim that its focus is entirely on Wikidata. Thus, the weak specialization is understandable. While it utilizes EL, the focus of the approach is still on knowledge base population. No training is necessary, which makes the approach suitable for a dynamic graph like Wikidata.

6.2.6. Rule-based approaches

Falcon 2.0 [100] is a fully linguistic approach and a transformation of Falcon 1.0 [99] to Wikidata. Falcon 2.0 was published in 2019, and its focus lies on short texts, especially questions. It links entities and relations jointly. Falcon 2.0 uses entity and relation labels as well as the triples themselves. The relations and entities are recognized by applying linguistic principles. The candidates are then generated by comparing mentions to the labels using the

Levenshtein distance. The ranking of the entities and relations is done by creating triples between the relations and entities and checking if the query is successful. The more successful the queries, the higher the candidate will be ranked. If no query is successful, the algorithm returns to the ER phase and splits some of the recognized entities again. As Falcon 2.0 is an extension of Falcon 1.0 from DBpedia to Wikidata, the usage of specific Wikidata characteristics is limited. Falcon 2.0 is tuned for EL on questions and short texts, as well as the English language and it was evaluated on the two QA datasets LC-QuAD 2.0 [27] and SimpleQuestions [13]. It is not generalizable to longer, more noisy, non-question texts. The used rules follow the structure of short questions. Hence, longer texts consisting of multiple sentences or non-questions are not supported. If the text is grammatically incorrect, the linguistic rules used to parse the utterance would fail. For example, linking Tweets would then be infeasible. As it is only based on rules, it is clearly independent of changes in the KG.

Tweeki [46] is an approach focusing on unsupervised EL over tweets. The ER is done by a pre-existing Entity Recognizer [40] which also tags the mentions. The candidates are generated by first calculating the link probability between Wikidata aliases over Wikipedia and then searching for the aliases in a dictionary. The ranking is done using the link probabilities while pruning all candidates that do not belong to the type provided by the Entity Recognizer. Tweeki was evaluated on the accompanied dataset TweekiGold, consisting of random annotated tweets. Additionally, it was tested on the Microposts 2016 [121] dataset and the datasets by Derczynski [25] which both also focus on shorter, noisy texts like tweets. The approach does not need to be trained, making it very suitable for linking entities in tweets. In this document type, often novel entities with minimal context exist. Regarding features of Wikidata, it uses label, alias and type information. Due to it being unsupervised, changes to the KG do not affect it.

6.3. Analysis

Many approaches include some form of language model or word embedding. This is expected as a large factor of entity linking encompasses the comparison of word-based information. And in that regard, language models like BERT [26] proved very performant in the last years. Furthermore, various language models rely on sub-word or character embeddings which also work on out-of-dictionary words. This is in contrast to regular word-embeddings, which can not cope with words never seen before. If graph information is part of the approach, the approaches either used graph embeddings, included some coherence score as a feature or created a neighborhood graph on the fly and optimized over it. Some approaches like OpenTapioca, Falcon 2.0 or Tweeki utilized more old-fashioned methods. They either employed classic ML together with some basic features or worked entirely rule-based.

6.3.1. Performance

Table 12 gives an overview of all available results for the approaches performing ER and EL. While results for the EL-only approaches exist, the used measures vary widely. Thus, it is very difficult to compare the approaches. To not withhold the results, they can still be found in the appendix in Table 16 with an accompanying discussion. We aim to fully recover this table and also extend Table 12 in future work.

The micro F_1 scores are given:

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r}$$

where p is the precision $p = \frac{tp}{tp+fp}$ and r is the recall $r = \frac{tp}{tp+fn}$. Here, tp is the number of true positives, fp is the number of false positives and fn is the number of false negatives over a ground truth. Micro F_1 means that the scores are calculated over all linked entity mentions and not separately for each document and then averaged. True positives are the correctly linked entity mentions, false positives incorrectly linked entities that do not occur in the set of valid entities and false negatives entities that occur in the set of valid entities but are not linked to [20]. The approaches were evaluated on many different datasets, which makes comparison very difficult. Additionally, many approaches are evaluated on datasets designed for knowledge graphs different from Wikidata and then mapped. Often, the approaches are evaluated on the same dataset but over different subsets, which complicates a comparison even more. The method by Perkins [86] was also evaluated on the Kensho Derived Wikimedia Dataset [56], but it

was only used to compare different variants of the designed approach and focused on different amounts of training data. Thus, inclusion in the evaluation table is not reasonable.

Inferring the utility of a Wikidata characteristic from the different approaches' F_1-measures is inconclusive due to the sparsity of results. For ER + EL approaches, most results were available for LC-QuAD 2.0. Yet, no conclusion can be drawn as many approaches were evaluated on different subsets of the dataset. Falcon 2.0 performs well, but it does not substantially rely on Wikidata characteristics. The performance is good as it is designed for simple questions that follow its rules very closely. Arjun performs well on T-REx by mainly using label information, but the number of methods tested on the T-REx dataset is too low to be conclusive. Besides that, PNEL and the approach by Huang et al. also achieve good results; both include a broader scope of Wikidata information in the form of labels, descriptions and graph structure. As HIPE challenge approaches are using Wikidata only marginally and the difference in performance depends more on the robustness against the OCR-introduced noise, comparing them is not providing information on the relevance of Wikidata characteristics.

6.3.2. Utilization of Wikidata characteristics

While some algorithms [78] do try to examine the challenges of Wikidata, like more noisy long entity labels, many fail to use most of the advantages of Wikidata's characteristics. If the approaches are using even more information than just the labels of entities and relations, they mostly only include simple n-hop triple information. Hyper-relational information like qualifiers is only used by OpenTapioca but still in a simple manner. This is surprising, as they can provide valuable additional information. As one can see in Fig. 8, around half of the statements on entities occurring in the LC-QuAD 2.0 dataset have one or more qualifiers. These percentages differ from the ones in all of Wikidata, but when entities are considered, appearing in realistic use cases like QA, qualifiers are much more abundant. Thus, dismissing the qualifier information might be critical. The inclusion of hyper-relational graph embeddings could improve the performance of many approaches already using non-hyper-relational ones. Rank information of statements might be useful to consider, but choosing the best one will probably often suffice.

Of all approaches, only two algorithms [8,53] use descriptions explicitly. Others incorporate them through triples too, but more on the side [77]. Descriptions can provide valuable context information and many items do have them; see Fig. 6d. Hedwig [60] claims to use descriptions but fails to describe how.

Two approaches [16,60] demonstrated the usefulness of the inherent multilingualism of Wikidata, notably in combination with Wikipedia.

As Wikidata is always changing, approaches robust against change are preferred. A reliance on transductive graph embeddings [8,18,86,105], which need to have all entities available during training, makes repeated training necessary. Alternatively, the used embeddings would need to be replaced with graph embeddings, which are efficiently updatable or inductive [3,6,22,38,45,110,122,123,128]. The rule-based approach Falcon 2.0 [100] is not affected by a developing knowledge graph but only usable for correctly-stated questions. Methods only working on text information [16,53,77,78,89] like labels, descriptions or aliases do not need to be updated if Wikidata changes,

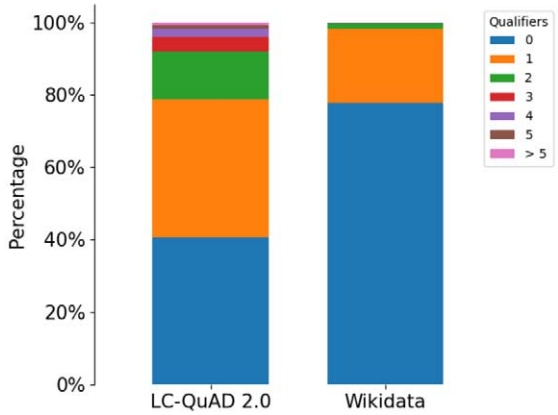


Fig. 8. Percentage of statements having the specified number of qualifiers for all LC-QuAD 2.0 and Wikidata entities.

Table 12
Results: ER + EL

	OpenTapioca [24]	Falcon 2.0 [100]	Arjun [78]	VCG [105]	KBPearl [69][1]	PNEL [8]	Huang et al. [53]	Boros et al. [15]	Provatorov et al. [89]	Labusch & Neudecker [65]	Hedwig [60]	Tweeki [46]
AIDA-CoNLL [51]	0.482 [24]	–	–	–	–	–	–	–	–	–	–	–
Microposts 2016 [121]	0.087 [24], 0.148 [46]	–	–	–	–	–	–	–	–	–	–	0.248 [46]
ISTEX-1000 [24]	0.87 [24]	–	–	–	–	–	–	–	–	–	–	–
RSS-500 [94]	0.335 [24]	–	–	–	–	–	–	–	–	–	–	–
LC-QuAD 2.0 [27]	0.301 [8]	0.445 [8]	–	0.47 [8]	–	0.589 [8][2]	–	–	–	–	–	–
LC-QuAD 2.0 [27][3]	0.25 [100]	0.68 [100]	–		–	–	–	–	–	–	–	–
LC-QuAD 2.0 [27][4]	–	0.320 [8]	–	–	–	0.629 [8][2]	–	–	–	–	–	–
Simple-Question	0.20 [8]	0.41 [8]	–	–	–	0.68 [8][5]	–	–	–	–	–	-
Simple-Question [13][6]	–	0.63 [100]	–	–	–	-	–	–	–	–	–	-
T-REx [33]	0.579 [78]	–	0.713 [78]	–	–	–	–	–	–	–	–	–
T-REx [33][7]	-	–	–	–	0.421 [69]	–	–	–	–	–	–	–
WebQSP [132]	–	–	0.730 [8,105]	–	-	0.712 [8][8]	0.780 [53]	–	–	–	–	–
CLEF HIPE 2020 [29]	–	–	–	–	–	–	–	0.531 [30][9]	0.300 [30][9]	0.141 [30][9]	–	–
TAC2017 [55]	–	–	–	–	–	-	–	–	–	–	0.582 [60]	
Graph-Questions [108]	–	–	0.442 [105]	–	–	–	–	–	–	–	–	–
QALD-7-WIKI [113]	–	–	–	–	0.679 [69]	–	–	–	–	–	–	–
NYT2018 [68,69]	–	–	–	–	0.575 [69]	–	–	–	–	–	–	–
ReVerb38 [69]	–	–	–	–	0.653 [69]	–	–	–	–	–	–	–
Knowledge Net [23]	–	–	–	–	0.384 [69]	–	–	–	–	–	–	–
CC-DBP [44]	–	–	–	–	0.499 [69]	–	–	–	–	–	–	–
TweekiGold [46]	0.291 [46]	–	–	–	–	–	–	–	–	–	–	0.65 [46]
Derczynski [25]	0.14 [46]	–	–	–	–	–	–	–	–	–	–	0.371 [46]

[1] NN model
[2] L model
[3] 1000 sampled questions from LC-QuAD 2.0
[4] LC-QuAD 2.0 test set used in KBPearl paper
[5] S model
[6] Probably evaluated on train and test set
[7] Evaluation on subset of T-REx data different to the subset used in Arjun paper
[8] W model
[9] Strict mention matching

only if the text type or the language itself does. This is demonstrated by the approach by Botha et al. [16] and the Wikification EL BLINK [127], which mainly use the BERT model and are able to link to entities never seen during training. If word-embeddings instead of sub-word embeddings are used, for example, GloVe [85] or word2vec [75], this advantage diminishes as new never-seen labels could not be interpreted. Nevertheless, the ability to support totally unseen new entities was only demonstrated for the approach by Botha et al [16]. The other approaches still need to be evaluated on the zero-shot EL task to be certain. For approaches [46,53,60] that rely on statistics over Wikipedia, new entities in Wikidata may sometimes not exist in Wikipedia to a satisfying degree. As a consequence, only a subset of all entities in Wikidata is supported. This also applies to the approaches by Boros et al. [15], and Labusch and Neudecker [65] which are mostly using Wikipedia information. Additionally, they are susceptible to changes in Wikipedia, especially specific statistics calculated over Wikipedia pages which have to be updated any time a new entity is added. Botha et al. [16] also mainly depend on Wikipedia and thus on the availability of the desired Wikidata entities in Wikipedia itself. Since the approach uses Wikipedia articles in multiple languages, it encompasses many more entities than the previous approaches that focus on Wikipedia. Botha et al.'s [16] approach was designed for the zero- and few-shot setting, it is quite robust against changes in the underlying knowledge graph.

Approaches relying on statistics [24,69] need to update them regularly, but this might be efficiently doable. Overall, the robustness against change might be negatively affected by static/transductive graph embeddings.

Answer – RQ 3. ***How do current Entity Linking approaches exploit the specific characteristics of Wikidata?***

The preceding summary and evaluation of the existing Wikidata Entity Linkers, together with Table 11 and the descriptions in Sections 6.1 and 6.2, provide an overview of all approaches with a focus on the incorporated Wikidata-characteristics.

Answer – RQ 4. ***Which Wikidata characteristics are unexploited by existing Entity Linking approaches?***

The most unexploited characteristics are the descriptions, the hyper-relational structure and the type information, as can be seen in Table 11. Nearly none of the found approaches exploited hyper-relational information in the form of qualifiers. And the one (i.e. OpenTapioca) using them did that in a simple way. As it is confirmed by the benchmarks that the inclusion of those can improve the performance of link prediction [37], this might also be the case for the task of EL. Furthermore, description information is still greatly underutilized. It can be a valuable piece of context information of an entity. Of course, it is not ideal as often the description can be short, especially for long-tail entities. A possible way to circumvent this challenge is the recent development of ***Abstract Wikipedia*** *[119], which will support the multilingual generation of descriptions in the future. While some approaches utilize type information, most use them to limit the set of valid entity candidates to instances of only a small subset of all types, namely person, organization and location. This is surprising as the non-included paper by Raiman and Raiman [90] shows that a fine-grained type system can heavily improve the entity linking performance. As mentioned in Section 4, rank information might also be used by not only including statements of the best rank but also of others. For example, in the case that statements exist which were valid at different points of time, including all could prove useful when linking documents of different ages. But such a special use case is not considered by any existing Wikidata EL approach. For some more ideas on how to include those characteristics in the future, please refer to section 8.2.*

6.4. Reproducibility

Not all approaches are available as a Web API or even as source code. An overview can be found in Table 13. The number of approaches for Wikidata having an accessible Web API is meager. While the code for some methods exists, this is the case for only half of them. The effort to set up different approaches also varies significantly due to missing instructions or data. Thus, we refrained from evaluating and filling the missing results for all the datasets in Tables 16 and 12. However, we seek to extend both tables in future work.

7. Related work

While there are multiple recent surveys on EL, none of those are specialized in analyzing EL on Wikidata.

Table 13
Availability of approaches

Approach	Code	Web API
OpenTapioca [24]	✓	✓
Falcon 2.0 [100]	✓	✓
Arjun [78]	✓	×
VCG [105]	✓	×
KBPearl [69]	×	×
PNEL [8]	✓	×
Mulang et al. [77]	✓	×
Perkins [86]	×	×
NED using DL on Graphs [18]	✓	×
Huang et al. [53]	×	×
Boros et al. [15]	×	×
Provatorov et al. [89]	×	×
Labusch and Neudecker [65]	✓	×
Botha et al. [16]	×	×
Hedwig [60]	×	×
Tweeki [46]	×	×

Table 14
Survey comparison

Survey	# Approaches	# Wikidata approaches	# Datasets	# Wikidata datasets
Sevgili et al. [101]	30	0	9	0
Al-Moslmi et al. [2]	39	0	17	0
Oliveira et al. [81]	36	0	32	0
This survey	16	16	21	11

The extensive survey by Sevgili et al. [101] is giving an overview of all neural approaches from 2015 to 2020. It compares 30 different approaches on nine different datasets. According to our criteria, none of the included approaches focuses on Wikidata. The survey also discusses the current state of the art of domain-independent and multi-lingual neural EL approaches. However, the influence of the underlying KG was not of concern to the authors. It is not described in detail how they found the considered approaches.

In the survey by Al-Moslmi et al. [2], the focus lies on ER and EL approaches over KGs in general. It considers approaches from 2014 to 2019. It gives an overview of the different approaches of ER, Entity Disambiguation, and EL. A distinction between Entity Disambiguation and EL is made, while our survey sees Entity Disambiguation as a part of EL. The roles of different domains, text types, or languages are discussed. The authors considered 89 different approaches and tools. Most approaches were designed for DBpedia or Wikipedia, some for Freebase or YAGO, and some to be KG-agnostic. Again, none focused on Wikidata. F_1 scores were gathered on 17 different datasets. Fifteen algorithms, for which an implementation or a WebAPI was available, were evaluated using GERBIL [95].

Another survey [81] examines recent approaches, which employ holistic strategies. Holism in the context of EL is defined as the usage of domain-specific inputs and metadata, joint ER-EL approaches, and collective disambiguation methods. Thirty-six research articles were found which had any holistic aspect – none of the designed approaches linked explicitly to Wikidata.

A comparison of the number of approaches and datasets included in the different surveys can be found in Table 14.

If we go further into the past, the existing surveys [71,102] are not considering Wikidata at all or only in a small amount as it is still a rather recent KG in comparison to the other established ones like DBpedia, Freebase or YAGO. For an overview of different KGs on the web, we refer the interested reader to the paper by Heist et al. [48].

No found survey focused on the differences of EL over different knowledge graphs, respectively, on the particularities of EL over Wikidata.

8. Discussion

8.1. Current approaches, datasets and their drawbacks

Approaches The number of algorithms using Wikidata is small; the number of algorithms using Wikidata solely is even smaller. Most algorithms employ labels and alias information contained in Wikidata. Some deep learning-based algorithms leverage the underlying graph structure, but the inclusion of that information is often superficial. The same information is also available in other KGs. Additional statement-specific information like qualifiers is used by only one algorithm (OpenTapioca), and even then, it only interprets qualifiers as extra edges to the item. Thus, there is no inclusion of the actual structure of a hyper-relation. Information like the descriptions of items that are providing valuable context information is also rarely used. Wikidata includes type information, but almost none of the existing algorithms utilize it to do more than to filter out entities that are not desired to link in general. An exception is perhaps Tweeki, though it only uses types during ER.

It seems that most of the authors developed approaches for Wikidata due to it being popular and up-to-date while not specifically utilizing its structure. With small adjustments, many would also work on any other KG. Besides the less-dedicated utilization of specific characteristics of Wikidata, it is also notable that there is no clear focus on one of the essential characteristics of Wikidata, continual growth. Many approaches use static graph embeddings, which need to be retrained if the KG changes. EL algorithms working on Wikidata, which are not usable on future versions, seem unintuitive. But there also exist some approaches which can handle change. They often rely on more extensive textual information, which is again challenging due to the limited amount of such data in Wikidata. Wikidata descriptions do exist, but only short paragraphs are provided, in general, insufficient to train a language model. To compensate, Wikipedia is included, which provides this textual information. It seems like Wikidata as the target KG with its language-agnostic identifiers and the easily connectable Wikipedia with its multilingual textual information are a great pair. But surprisingly, most methods do use either Wikipedia or Wikidata. A combination happens rarely but seems very fruitful, as can be seen via the performance of the multilingual EL by Botha et al. [16]. Though even this approach still uses Wikidata only sparsely.

None of the investigated approaches' authors tried to examine the performance between different versions of Wikidata. Since continuous evolution is a central characteristic of Wikidata, a temporal analysis would be reasonable. As we are confronted with a fast-growing ocean of knowledge, taking into account the change of Wikidata and hence developing approaches that are robust against that change will undoubtedly be useful for numerous applications and their users.

This survey aimed to identify the extent to which the current state of the art in Wikidata EL is utilizing the characteristics of Wikidata. As only a few are using more information than on other established KGs, there is still much potential for future research.

Datasets Only a limited number of datasets were created entirely with Wikidata in mind exist. Many datasets used are still only mapped versions of datasets created for other knowledge graphs. Multilingualism is present so far that some datasets contain documents in different languages. However, only different documents for different languages are available. Having the same documents in multiple languages would be more helpful for an evaluation of multilingual Entity Linkers. The fact that the Wikidata is ever-changing is also not genuinely considered in any datasets. Always providing the dump version on which the dataset was created is advisable. A big advantage for the community is that datasets from very different domains like news, forums, research, tweets exist. The utterances can also vary from shorter texts with only a few entities to large documents with many entities. The difficulty of the datasets significantly differs in the ambiguity of the entity mentions. The datasets also differ in quality. Some were automatically created and others annotated manually by experts. There are no unanimously agreed-upon datasets used for Wikidata EL. Of course, a single dataset can not exist as different domains and text types make different approaches, and hence datasets necessary.

8.2. Future research avenues

In general, Wikidata EL could be improved by including the following aspects:

Hyper-relational statements The qualifier and rank information of Wikidata could be suitable to do EL on time-sensitive utterances [1]. The problem revolves around utterances that talk about entities from different time points and spans and thus, the referred entity can significantly diverge. The usefulness of other characteristics of Wikidata, e.g., references, may be limited but could make EL more challenging due to the inclusion of contradictory information. Therefore, research into the consequences and solutions of conflicting information would be advisable. Another possibility would be to directly include the qualifier information via the KG embeddings. For example, the StarE [37] embedding includes qualifiers directly during training. It performs superior over regular embeddings on the task of link prediction if enough statements have qualifiers assigned. This is promising but whether this directly applies to EL approaches, which use such embeddings, has to be evaluated.

More extensive type information While type information is incorporated by some linkers, it is generally done to simply limit the candidate space to the three main types: location, organization and person. But Raiman and Raiman [90] showed that a more extensive system of types proves very effective on the task of EL. If an adequate typing system is chosen and the correct type of an entity mention is available, an entity linker can achieve a near-perfect performance. Especially as Wikidata has a much more fine-grained and noisy type system than other KGs, evaluating the performance of entity linkers, which incorporate types, is of interest. While most approaches use types directly to limit the candidate space, incorporating them indirectly via type-aware [135] or hierarchy-sensitive embeddings [7,19,79] might also prove useful for EL. But note that the incorporation of type information heavily depends on the performance of the type classifier, and the difficulty of the type classification task again depends on the type system. Nevertheless, an improved type classification would directly benefit type-utilizing entity linkers.

Inductive or efficiently trainable knowledge graph embeddings To reiterate, due to the fast rate of change of Wikidata, approaches are necessary, which are more robust to such a dynamic KG. Continuously retraining transductive embeddings is intractable, so more sophisticated methods like inductive or efficiently retrievable graph embeddings are a necessity [3,6,22,38,45,110,122,123,128]. For example, the embedding by Albooyeh et al. [3] can be employed, which can handle out-of-sample entities. These are entities, which were not available at training time, but are connected to entities, which were existing. To go even further, NodePiece [38], the KG-embedding counterpart of sub-word embeddings like BERT, works by relying on only a small subset of anchor nodes and all relations in the KG. While it uses a fraction of all nodes, it still is able to achieve performance competitive with transductive embeddings on the task of link prediction. By being independent of most nodes in a KG, one can include new entities (in the form of nodes) without having to retrain. As an alternative, standard continual learning approaches could be employed to learn new data while being robust against catastrophic forgetting. An examination of the performance of popular techniques in the context of KG embeddings can be found in the paper by Daruna et al. [22].

Item label and description information in multiple languages for multilingual EL Multilingual or cross-lingual EL is already tackled with Wikidata but currently mainly by depending on Wikipedia. Using the available multilingual label/description information in a structured form together with the rich textual information in Wikipedia could move the field forward. The approach by Botha et al. [16], which could be seen as an extension of BLINK [127], performs very well on the task of cross- and multilingual EL. For example, the approach by Mulang et al. [77], which fully relies on label information, could be extended in a similar way as BLINK was extended. Instead of only using labels (of items and properties) in the English language, training the model directly in multiple languages could prove effective. Additionally, multilingual description information might be used too. We are convinced that also investigations into the linking of long-tail entities are needed.

It seems like there exist no commonly agreed-on Wikidata EL datasets, as shown by a large number of different datasets the approaches were tested on. Such datasets should try to represent the challenges of Wikidata like the time-variance, contradictory triple information, noisy labels, and multilingualism.

Acknowledgements

We acknowledge the support of the EU project TAILOR (GA 952215), the Federal Ministry for Economic Affairs and Energy (BMWi) project SPEAKER (FKZ 01MK20011A), the German Federal Ministry of Education and

Research (BMBF) projects and excellence clusters ML2R (FKZ 01 15 18038 A/B/C), MLwin (01S18050 D/F), ScaDS.AI (01/S18026A) as well as the Fraunhofer Zukunftsstiftung project JOSEPH. The authors also acknowledge the financial support by the Federal Ministry for Economic Affairs and Energy of Germany in the project CoyPu (project number 01MK21007G).

Appendix A. KG-agnostic entity linkers

AGDISTIS [114] is an EL approach expecting already marked entity mentions. It expects a KG dump available in the Turtle format [10]. For candidate generation, first, an index is created which contains all available entities and their labels. They are extracted from the available Turtle dump. The input entity mention is first normalized by reducing plural and genitive forms and removing common affixes. Furthermore, if an entity mention consists of a substring of a preceding entity mention, the succeeding one is directly mapped to the preceding one. Additionally, the space of possible candidates can be limited by configuration. Usually, the candidate space is reduced to organizations, persons and locations. The candidates are then searched for over the index by comparing the reduced entity mention with the labels in the index using trigram similarity. No candidates are included, which contain time information inside the label. After gathering all candidates of all entity mentions in the utterance, the candidates are ranked by building a temporary graph. Starting with the candidates as the initial nodes, the graph is expanded breadth-first by adding the adjacent nodes and the edges in-between. It is done to some previously set depth. This results in a partly connected graph containing all candidates. Then the HITS-algorithm [61] is run and the most authoritative candidate nodes are chosen per entity mention. Thus, the approach is performing a global entity coherence optimization. The approach uses label and alias information for building the index. Type information can be used to restrict the candidate space and the KG structure is utilized during the candidate ranking.

MAG MAG [76] is a multilingual extension of AGDISTIS. Again, no ER is performed. The same label index as used in AGDISTIS is employed. Besides that, the following additional indices were created:

- A person index, containing the person names and the variations in different languages
- A rare references index containing textual descriptions of entities
- An acronym index based on the commercial STANDS4[22] data
- A context index containing semantic embeddings of Concise Bounded Description[23]

During candidate generation, it is first checked if the entity mention corresponds to an acronym. If it is one, no further preprocessing is done. If not, the entity mention is normalized by removing special characters, changing the casing and splitting camel-cased words. After preprocessing, the candidates are searched by first checking for exact matches, then searching via trigram similarity. If this still did not produce any candidates, the entity mention is stemmed and the search is repeated. If a mention is an acronym, the candidate list is expanded with the corresponding entities. Then, more candidates are searched by taking an entity mention and the set of all entity mentions in the utterance. Those are used to build a tf-idf search query over the context index. All returned candidates are then first filtered by trigram similarity between entity mention and candidate. A second filtering is applied by counting the number of direct connections between the remaining candidates and the candidates of the other entity mentions. The candidates with too few links are pruned away. All the candidates of the entity mention are then sorted by their popularity (calculated via PageRank [83]) and the top 100 are returned. Then, the entities are disambiguated in nearly the same way as done by AGDISTIS. The only difference is the option to use PageRank instead of HITS to rank the final candidates. Additionally to the properties already used by AGDISTIS, item descriptions are incorporated via the context index.

DoSeR [137] also expects already marked entity mentions. The linker focuses being to link to multiple knowledge graphs simultaneously. Here, they support RDF-based KGs and entity-annotated document (EAD) KGs (e.g., Wikipedia). The KGs are split into core and optional KGs. Core KGs contain the entities to which one wants to link.

[22] http://www.abbreviations.com/
[23] https://www.w3.org/Submission/CBD/

Table 15
Comparison between the utilized Wikidata characteristics of each KG-agnostic approach

Approach	Labels/Aliases	Descriptions	Knowledge graph structure	Hyper-relational structure	Types	Additional information
AGDISTIS [114]	✓	×	✓	×	✓	
MAG [76]	✓	✓	✓	×	✓	STANDS4[22]
DoSeR [137]	✓	×	✓	×	✓	

Table 16
Results: EL-only

	Mulang et al. [77]	LSH-ELMo model [86]	NED using DL on Graphs [18][1]	Botha et al. [16]
AIDA-CoNLL [51]	0.9494 [77][2,3]	0.73 [86]	–	–
ISTEX-1000 [24]	0.9261 [77][4]	–	–	–
Wikidata-Disamb [18]	0.9235 [77][5]	–	0.916 [18]	–
Mewsli-9 [16]	–	–	–	0.91 [16][6]

[1]Model with best result
[2]Accuracy instead of F_1
[3]DCA-SL used
[4]XLNet used
[5]Roberta used
[6]Recall instead of F_1

Optional KGs complement the core KGs with additional data. First, an index is created which includes the entities of all core KGs. In the index, the labels or surface forms, a semantic embedding, and each entity's prior popularity are stored. The semantic embeddings are computed by using Word2Vec. For EAD-KGs, the different documents are taken and all words, which are not pointing to entities, are removed. All remaining words are replaced with the corresponding entity identifier. These sequences are then used to train the embeddings. For RDF-KGs, a Random Walk is performed over the graph and the resulting sequences are used to train the embeddings. The succeeding node is chosen with a probability corresponding to the reciprocal of the number of edges it got. The same probability is used to sometimes jump to another arbitrary node in the graph. The prior probability is calculated by either using the number of incoming/outgoing edges in the RDF-KG or the number of annotations that point to the entity in the EAD KG. If type information is available, the entity space can be limited here too. First, candidates are generated by searching for exact matches and then the AGDISTIS candidate generation is used to find more candidates. The candidates are disambiguated, similar to the way AGDISTIS and MAG are doing it. First, a graph is built though not a complete graph but a K-partite graph where K is the number of all entity mentions. Edges exist only between candidates of different entities. Using the complete graph resulted in a loss of performance. After the graph is created, PageRank is done to score the different entities coherently. The edge weights correspond to the (normalized) cosine similarity of the semantic embeddings of the two connected entities. Additionally, at any point during PageRank computation, it is possible to jump to an arbitrary node with a certain probability. This probability depends on the prior popularity of the entity. It uses label information, the knowledge graph structure and type information (if desired).

Appendix B. EL-only results and discussion

The results for EL-only approaches can be found in Table 16. AIDA-CoNLL results are available for three of the four approaches, but the results for one is the accuracy instead of the F_1-measures. The available labels for each item and property make language-model-based approaches possible that perform quite well [77]. No approaches are available to compare to the one by Botha et al. [16], but the result demonstrates the promising performance of multilingual EL with Wikidata as the target KG.

Table 17
Links to datasets

Dataset	Links
T-REx [33]	https://hadyelsahar.github.io/t-rex
NYT2018 [68,69]	not found
ISTEX-1000 [24]	https://github.com/wetneb/opentapioca/blob/master/data
LC-QuAD 2.0 [27]	https://github.com/AskNowQA/LC-QuAD2.0/tree/master/dataset
Knowledge Net [23]	https://github.com/diffbot/knowledge-net/tree/master/dataset
KORE50DYWC [80]	https://www.aifb.kit.edu/web/KORE_50%5EDYWC
Kensho Derived Wikimedia Dataset [56]	https://www.kaggle.com/kenshoresearch/kensho-derived-wikimedia-data
CLEF HIPE 2020 [29]	https://github.com/impresso/CLEF-HIPE-2020/tree/master/data
Mewsli-9 [16]	https://metatext.io/datasets/mewsli-9-
TweekiData [46]	https://github.com/ucinlp/tweeki/tree/main/data/Tweeki_data
TweekiGold [46]	https://github.com/ucinlp/tweeki/tree/main/data/Tweeki_gold

References

[1] P. Agarwal, J. Strötgen, L.D. Corro, J. Hoffart and G. Weikum, diaNED: Time-aware named entity disambiguation for diachronic corpora, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, Melbourne, Australia, July 15–20, 2018, S. Papers, I. Gurevych and Y. Miyao, eds, Vol. 2, Association for Computational Linguistics, 2018, pp. 686–693, https://www.aclweb.org/anthology/P18-2109/. doi:10.18653/v1/P18-2109.

[2] T. Al-Moslmi, M.G. Ocaña, A.L. Opdahl and C. Veres, Named entity extraction for knowledge graphs: A literature overview, *IEEE Access* **8** (2020), 32862–32881. doi:10.1109/ACCESS.2020.2973928.

[3] M. Albooyeh, R. Goel and S.M. Kazemi, Out-of-sample representation learning for knowledge graphs, in: *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event*, 16–20 November, 2020, T. Cohn, Y. He and Y. Liu, eds, Findings of ACL, Vol. EMNLP 2020, Association for Computational Linguistics, 2020, pp. 2657–2666. doi:10.18653/v1/2020.findings-emnlp.241.

[4] P.D. Almeida, J.G. Rocha, A. Ballatore and A. Zipf, Where the streets have known names, in: *Proceedings, Part IV, Computational Science and Its Applications – ICCSA 2016 – 16th International Conference*, Beijing, China, July 4–7, 2016, O. Gervasi, B. Murgante, S. Misra, A.M.A.C. Rocha, C.M. Torre, D. Taniar, B.O. Apduhan, E.N. Stankova and S. Wang, eds, Lecture Notes in Computer Science, Vol. 9789, Springer, 2016, pp. 1–12. doi:10.1007/978-3-319-42089-9_1.

[5] G. Angeli, M.J.J. Premkumar and C.D. Manning, Leveraging linguistic structure for open domain information extraction, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 1: Long Papers*, Beijing, China, July 26–31, 2015, The Association for Computer Linguistics, 2015, pp. 344–354. doi:10.3115/v1/p15-1034.

[6] J. Baek, D.B. Lee and S.J. Hwang, Learning to extrapolate knowledge: Transductive few-shot out-of-graph link prediction, in: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual*, December 6–12, 2020, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, eds, 2020, https://proceedings.neurips.cc/paper/2020/hash/0663a4ddceacb40b095eda264a85f15c-Abstract.html.

[7] I. Balazevic, C. Allen and T.M. Hospedales, Multi-relational Poincaré graph embeddings, in: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, 2019, NeurIPS 2019*, Vancouver, BC, Canada, December 8–14, 2019, H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox and R. Garnett, eds, 2019, pp. 4465–4475, https://proceedings.neurips.cc/paper/2019/hash/f8b932c70d0b2e6bf071729a4fa68dfc-Abstract.html.

[8] D. Banerjee, D. Chaudhuri, M. Dubey and J. Lehmann, PNEL: Pointer network based end-to-end entity linking over knowledge graphs, in: *Proceedings, Part I, The Semantic Web – ISWC 2020–19th International Semantic Web Conference*, Athens, Greece, November 2–6, 2020, J.Z. Pan, V.A.M. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal, eds, Lecture Notes in Computer Science, Vol. 12506, Springer, 2020, pp. 21–38. doi:10.1007/978-3-030-62419-4_2.

[9] A. Bastos, A. Nadgeri, K. Singh, I.O. Mulang, S. Shekarpour, J. Hoffart and M. Kaul, RECON: Relation extraction using knowledge graph context in a graph neural network, in: *WWW'21: The Web Conference 2021, Virtual Event*, Ljubljana, Slovenia, April 19–23, 2021, J. Leskovec, M. Grobelnik, M. Najork, J. Tang and L. Zia, eds, ACM/IW3C2, 2021, pp. 1673–1685. doi:10.1145/3442381.3449917.

[10] D. Beckett, T. Berners-Lee, E. Prud'hommeaux and G. Carothers, in: *RDF 1.1 Turtle, World Wide Web Consortium*, 2014, pp. 18–31.

[11] F. Blog, *Introducing the Freebase RDF Service*, 2008, https://web.archive.org/web/20120516075431/http://blog.freebase.com/2008/10/30/introducing_the_rdf_service/.

[12] K.D. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, Freebase: A collaboratively created graph database for structuring human knowledge, in: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008*, Vancouver, BC, Canada, June 10–12, 2008, J.T. Wang, ed., ACM, 2008, pp. 1247–1250. doi:10.1145/1376616.1376746.

[13] A. Bordes, N. Usunier, S. Chopra and J. Weston, Large-scale Simple Question Answering with Memory Networks, 2015, CoRR, http://arxiv.org/abs/1506.02075 arXiv:1506.02075.

[14] A. Bordes, N. Usunier, A. García-Durán, J. Weston and O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Proceedings of a Meeting Held December 5–8, 2013*, Lake Tahoe, Nevada, United States, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 2787–2795, http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.

[15] E. Boros, E.L. Pontes, L.A. Cabrera-Diego, A. Hamdi, J.G. Moreno, N. Sidère and A. Doucet, Robust named entity recognition and linking on historical multilingual documents, in: *Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece, September 22–25, 2020, L. Cappellato, C. Eickhoff, N. Ferro and A. Névéol, eds, CEUR Workshop Proceedings, Vols 2696, CEUR-WS.org, 2020, http://ceur-ws.org/Vol-2696/paper_171.pdf.

[16] J.A. Botha, Z. Shan and D. Gillick, Entity linking in 100 languages, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, November 16–20, 2020, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 7833–7845. doi:10.18653/v1/2020.emnlp-main.630.

[17] M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov et al., Deeppavlov: Open-source library for dialogue systems, in: *Proceedings of ACL 2018, System Demonstrations*, 2018, pp. 122–127. doi:10.18653/v1/P18-4021.

[18] A. Cetoli, M. Akbari, S. Bragaglia, A.D. O'Harney and M. Sloan, Named Entity Disambiguation using Deep Learning on Graphs, 2018, CoRR, http://arxiv.org/abs/1810.09164 arXiv:1810.09164.

[19] I. Chami, A. Wolf, D. Juan, F. Sala, S. Ravi and C. Ré, Low-dimensional hyperbolic knowledge graph embeddings, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*, July 5–10, 2020, D. Jurafsky, J. Chai, N. Schluter and J.R. Tetreault, eds, Association for Computational Linguistics, 2020, pp. 6901–6914. doi:10.18653/v1/2020.acl-main.617.

[20] M. Cornolti, P. Ferragina and M. Ciaramita, A framework for benchmarking entity-annotation systems, in: *22nd International World Wide Web Conference, WWW'13*, Rio de Janeiro, Brazil, May 13–17, 2013, D. Schwabe, V.A.F. Almeida, H. Glaser, R. Baeza-Yates and S.B. Moon, eds, International World Wide Web Conferences Steering Committee /, ACM, 2013, pp. 249–260. doi:10.1145/2488388.2488411.

[21] L.D. Corro and R. Gemulla, ClausIE: Clause-based open information extraction, in: *22nd International World Wide Web Conference, WWW'13*, Rio de Janeiro, Brazil, May 13–17, 2013, D. Schwabe, V.A.F. Almeida, H. Glaser, R. Baeza-Yates and S.B. Moon, eds, International World Wide Web Conferences Steering Committee /, ACM, 2013, pp. 355–366. doi:10.1145/2488388.2488420.

[22] A.A. Daruna, M. Gupta, M. Sridharan and S. Chernova, Continual learning of knowledge graph embeddings, *IEEE Robotics Autom. Lett.* **6**(2) (2021), 1128–1135. doi:10.1109/LRA.2021.3056071.

[23] F. de Sá Mesquita, M. Cannaviccio, J. Schmidek, P. Mirza and D. Barbosa, KnowledgeNet: A benchmark dataset for knowledge base population, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, Hong Kong, China, November 3–7, 2019, K. Inui, J. Jiang, V. Ng and X. Wan, eds, Association for Computational Linguistics, 2019, pp. 749–758. doi:10.18653/v1/D19-1069.

[24] A. Delpeuch, OpenTapioca: Lightweight entity linking for wikidata, in: *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) Co-Located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference*, November 2–6, 2020, L. Kaffee, O. Tifrea-Marciuska, E. Simperl and D. Vrandecic, eds, CEUR Workshop Proceedings, Vols 2773, CEUR-WS.org, 2020, http://ceur-ws.org/Vol-2773/paper-02.pdf.

[25] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak and K. Bontcheva, Analysis of named entity recognition and linking for tweets, *Inf. Process. Manag.* **51**(2) (2015), 32–49. doi:10.1016/j.ipm.2014.10.006.

[26] J. Devlin, M. Chang, K. Lee and K. Toutanova, 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: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, June 2–7, 2019, J. Burstein, C. Doran and T. Solorio, eds, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[27] M. Dubey, D. Banerjee, A. Abdelkawi and J. Lehmann, LC-QuAD 2.0: A large dataset for complex question answering over wikidata and DBpedia, in: *Proceedings, Part II, The Semantic Web – ISWC 2019 – 18th International Semantic Web Conference*, Auckland, New Zealand, October 26–30, 2019, C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I.F. Cruz, A. Hogan, J. Song, M. Lefrançois and F. Gandon, eds, Lecture Notes in Computer Science, Vol. 11779, Springer, 2019, pp. 69–78. doi:10.1007/978-3-030-30796-7_5.

[28] L. Ehrlinger and W. Wöß, Towards a definition of knowledge graphs, in: *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems – SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) Co-Located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016)*, Leipzig, Germany, September 12–15, 2016, M. Martin, M. Cuquet and E. Folmer, eds, CEUR Workshop Proceedings, Vol. 1695, CEUR-WS.org, http://ceur-ws.org/Vol-1695/paper4.pdf.

[29] M. Ehrmann, M. Romanello, A. Flückiger and S. Clematide, Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers, in: *CLEF*, 2020a.

[30] M. Ehrmann, M. Romanello, A. Flückiger and S. Clematide, Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers, in: *CLEF*, 2020b.

[31] C.B. El Vaigh, G. Le Noé-Bienvenu, G. Gravier and P. Sébillot, *IRISA System for Entity Detection and Linking at CLEF HIPE 2020, in: CEUR Workshop Proceedings*, 2020.

[32] R. Ellgren, Exploring Emerging Entities and Named Entity Disambiguation in News Articles, Master's thesis, Linköping University, 2020.

[33] H. ElSahar, P. Vougiouklis, A. Remaci, C. Gravier, J.S. Hare, F. Laforest and E. Simperl, T-REx: A large scale alignment of natural language with knowledge base triples, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, Miyazaki, Japan, May 7–12, 2018, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga, eds, European Language Resources Association (ELRA), 2018, http://www.lrec-conf.org/proceedings/lrec2018/summaries/632.html.

[34] A. Fader, S. Soderland and O. Etzioni, Identifying relations for open information extraction, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL*, Edinburgh, UK, 27–31 July 2011, John McIntyre Conference Centre, 2011, pp. 1535–1545, https://www.aclweb.org/anthology/D11-1142/.

[35] M. Färber, F. Bartscherer, C. Menne and A. Rettinger, Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO, *Semantic Web* **9**(1) (2018), 77–129. doi:10.3233/SW-170275.

[36] W. Foundation, Wikistats, 2020-10-09, https://stats.wikimedia.org/#/metrics/wikidata.org.

[37] M. Galkin, P. Trivedi, G. Maheshwari, R. Usbeck and J. Lehmann, Message passing for hyper-relational knowledge graphs, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, November 16–20, 2020, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 7346–7359. doi:10.18653/v1/2020.emnlp-main.596.

[38] M. Galkin, J. Wu, E. Denis and W.L. Hamilton, NodePiece: Compositional and Parameter-Efficient Representations of Large Knowledge Graphs.

[39] O. Ganea and T. Hofmann, Deep joint entity disambiguation with local neural attention, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark, September 9–11, 2017, M. Palmer, R. Hwa and S. Riedel, eds, Association for Computational Linguistics, 2017, pp. 2619–2629. doi:10.18653/v1/d17-1277.

[40] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N.F. Liu, M.E. Peters, M. Schmitz and L. Zettlemoyer, *AllenNLP: A Deep Semantic Natural Language Processing Platform, CoRR abs/1803.07640*, 2018, http://arxiv.org/abs/1803.07640.

[41] K. Gashteovski, R. Gemulla and L.D. Corro, MinIE: Minimizing facts in open information extraction, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark, September 9–11, 2017, M. Palmer, R. Hwa and S. Riedel, eds, Association for Computational Linguistics, 2017, pp. 2630–2640. doi:10.18653/v1/d17-1278.

[42] D. Gillick, S. Kulkarni, L. Lansing, A. Presta, J. Baldridge, E. Ie and D. García-Olano, Learning dense representations for entity retrieval, in: *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019*, Hong Kong, China, November 3–4, 2019, M. Bansal and A. Villavicencio, eds, Association for Computational Linguistics, 2019, pp. 528–537. doi:10.18653/v1/K19-1049.

[43] A. Gionis, P. Indyk and R. Motwani, Similarity search in high dimensions via hashing, in: *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, UK, September 7–10, 1999, M.P. Atkinson, M.E. Orlowska, P. Valduriez, S.B. Zdonik and M.L. Brodie, eds, Morgan Kaufmann, 1999, pp. 518–529, http://www.vldb.org/conf/1999/P49.pdf.

[44] M.R. Glass and A. Gliozzo, A dataset for web-scale knowledge base population, in: *The Semantic Web – 15th International Conference, ESWC 2018, Proceedings*, Heraklion, Crete, Greece, June 3–7, 2018, A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai and M. Alam, eds, Lecture Notes in Computer Science, Vol. 10843, Springer, 2018, pp. 256–271. doi:10.1007/978-3-319-93417-4_17.

[45] T. Hamaguchi, H. Oiwa, M. Shimbo and Y. Matsumoto, Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach, in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, Melbourne, Australia, August 19–25, 2017, C. Sierra, ed., ijcai.org, 2017, pp. 1802–1808. doi:10.24963/ijcai.2017/250.

[46] B. Haradizadeh and S. Singh, Tweeki: Linking Named Entities on Twitter to a Knowledge Graph, 2020, in: Workshop on Noisy User-generated Text (W-NUT).

[47] S. Heindorf, M. Potthast, B. Stein and G. Engels, Vandalism detection in Wikidata, in: *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016*, Indianapolis, IN, USA, October 24–28, 2016, S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li and P. Sondhi, eds, ACM, 2016, pp. 327–336. doi:10.1145/2983323.2983740.

[48] N. Heist, S. Hertling, D. Ringler and H. Paulheim, Knowledge graphs on the web – an overview, in: *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, I. Tiddi, F. Lécué and P. Hitzler, eds, Studies on the Semantic Web, Vol. 47, IOS Press, 2020, pp. 3–22. doi:10.3233/SSW200009.

[49] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.* **9**(8) (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[50] J. Hoffart, Y. Altun and G. Weikum, Discovering emerging entities with ambiguous names, in: *23rd International World Wide Web Conference*, WWW'14, Seoul, Republic of Korea, April 7–11, 2014, C. Chung, A.Z. Broder, K. Shim and T. Suel, eds, ACM, 2014, pp. 385–396. doi:10.1145/2566486.2568003.

[51] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater and G. Weikum, Robust disambiguation of named entities in text, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011*, 27–31 July 2011, A Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, John McIntyre Conference Centre, Edinburgh, UK, 2011, pp. 782–792, https://www.aclweb.org/anthology/D11-1072/.

[52] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J.E.L. Gayo, S. Kirrane, S. Neumaier, A. Polleres et al., Knowledge graphs, 2020, Preprint arXiv:2003.02320.

[53] B. Huang, H. Wang, T. Wang, Y. Liu and Y. Liu, Entity Linking for Short Text Using Structured Knowledge Graph via Multi-grained Text Matching, *Amazon Science* (2020).

[54] M. Jarke, B. Neumann, Y. Vassiliou and W. Wahlster, KBMS requirements of knowledge-based systems, in: *Foundations of Knowledge Base Management*, Springer, 1989, pp. 381–394. doi:10.1007/978-3-642-83397-7_17.

[55] H. Ji, X. Pan, B. Zhang, J. Nothman, J. Mayfield, P. McNamee and C. Costello, Overview of TAC-KBP2017 13 languages entity discovery and linking, in: *Proceedings of the 2017 Text Analysis Conference, TAC 2017*, Gaithersburg, Maryland, USA, November 13–14, 2017, NIST, 2017, https://tac.nist.gov/publications/2017/additional.papers/TAC2017.KBP_Entity_Discovery_and_Linking_overview.proceedings.pdf.

[56] Kensho R&D group, Kensho Derived Wikimedia Dataset, 2020, https://www.kaggle.com/kenshoresearch/kensho-derived-wikimedia-data.

[57] B. Kitchenham, *Procedures for Performing Systematic Reviews*, Vol. 33, Keele University, Keele, UK, 2004, pp. 1–26.

[58] M. Klang, F. Dib and P. Nugues, Overview of the Ugglan Entity Discovery and Linking System, 2019, CoRR, http://arxiv.org/abs/1903.05498 arXiv:1903.05498.

[59] M. Klang and P. Nugues, Named entity disambiguation in a question answering system, in: *TheFifth Swedish Language Technology Conference (SLTC2014)*, 2014.

[60] M. Klang and P. Nugues, Hedwig: A named entity linker, in: *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020*, Marseille, France, May 11–16, 2020, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association, 2020, pp. 4501–4508, https://www.aclweb.org/anthology/2020.lrec-1.554/.

[61] J.M. Kleinberg, Hubs, authorities, and communities, *ACM Comput. Surv.* **31**(4es) (1999), 5. doi:10.1145/345966.345982.

[62] J. Klie, R.E. de Castilho and I. Gurevych, From zero to hero: Human-in-the-loop entity linking in low resource domains, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*, July 5–10, 2020, D. Jurafsky, J. Chai, N. Schluter and J.R. Tetreault, eds, Association for Computational Linguistics, 2020, pp. 6982–6993, https://www.aclweb.org/anthology/2020.acl-main.624/. doi:10.18653/v1/2020.acl-main.624.

[63] N. Kolitsas, O. Ganea and T. Hofmann, End-to-end neural entity linking, in: *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018*, Brussels, Belgium, October 31–November 1, 2018, A. Korhonen and I. Titov, eds, Association for Computational Linguistics, 2018, pp. 519–529. doi:10.18653/v1/k18-1050.

[64] T. Kristanti and L. Romary, DeLFT and entity-fishing: Tools for CLEF HIPE 2020 shared task, in: *CLEF 2020-Conference and Labs of the Evaluation Forum*, Vol. 2696, CEUR, 2020.

[65] K. Labusch and C. Neudecker, *Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT, Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*, 2020.

[66] P. Le and I. Titov, Improving entity linking by modeling latent relations between mentions, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, Melbourne, Australia, July 15–20, 2018, I. Gurevych and Y. Miyao, eds, Association for Computational Linguistics, 2018, pp. 1595–1604, https://aclanthology.org/P18-1148/. doi:10.18653/v1/P18-1148.

[67] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer and C. Bizer, DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia, *Semantic Web* **6**(2) (2015), 167–195. doi:10.3233/SW-140134.

[68] X. Lin and L. Chen, Canonicalization of open knowledge bases with side information from the source text, in: *35th IEEE International Conference on Data Engineering, ICDE 2019*, Macao, China, April 8–11, 2019, IEEE, 2019, pp. 950–961. doi:10.1109/ICDE.2019.00089.

[69] X. Lin, H. Li, H. Xin, Z. Li and L. Chen, KBPearl: A knowledge base population system supported by joint entity and relation linking, *Proc. VLDB Endow.* **13**(7) (2020), 1035–1049, http://www.vldb.org/pvldb/vol13/p1035-lin.pdf. doi:10.14778/3384345.3384352.

[70] Y. Lin, Z. Liu and M. Sun, Neural relation extraction with multi-lingual attention, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers*, Vancouver, Canada, July 30–August 4, 2017, R. Barzilay and M. Kan, eds, Association for Computational Linguistics, 2017, pp. 34–43. doi:10.18653/v1/P17-1004.

[71] X. Ling, S. Singh and D.S. Weld, Design challenges for entity linking, *Trans. Assoc. Comput. Linguistics* **3** (2015), 315–328, https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/528. doi:10.1162/tacl_a_00141.

[72] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019, CoRR, http://arxiv.org/abs/1907.11692 arXiv:1907.11692.

[73] L. Logeswaran, M. Chang, K. Lee, K. Toutanova, J. Devlin and H. Lee, Zero-shot entity linking by reading entity descriptions, in: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, Florence, Italy, July 28–August 2, 2019, A. Korhonen, D.R. Traum and L. Màrquez, eds, Association for Computational Linguistics, 2019, pp. 3449–3460. doi:10.18653/v1/p19-1335.

[74] M. Manske, Wikidata Stats, pp. 2020–07–2020-20, https://wikidata-todo.toolforge.org/stats.php.

[75] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, in: *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA, May 2–4, 2013, Y. Bengio and Y. LeCun, eds, http://arxiv.org/abs/1301.3781.

[76] D. Moussallem, R. Usbeck, M. Röder and A.N. Ngomo, MAG: A multilingual, knowledge-base agnostic and deterministic entity linking approach, in: *Proceedings of the Knowledge Capture Conference, K-CAP 2017*, Austin, TX, USA, December 4–6, 2017, Ó. Corcho, K. Janowicz, G. Rizzo, I. Tiddi and D. Garijo, eds, ACM, 2017, pp. 9:1–9:8. doi:10.1145/3148011.3148024.

[77] I.O. Mulang, K. Singh, C. Prabhu, A. Nadgeri, J. Hoffart and J. Lehmann, Evaluating the impact of knowledge graph context on entity disambiguation models, in: *CIKM'20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event*, Ireland, October 19–23, 2020, M. d'Aquin, S. Dietze, C. Hauff, E. Curry and P. Cudré-Mauroux, eds, ACM, 2020, pp. 2157–2160. doi:10.1145/3340531.3412159.

[78] I.O. Mulang, K. Singh, A. Vyas, S. Shekarpour, M. Vidal and S. Auer, Encoding knowledge graph entity aliases in attentive neural network for wikidata entity linking, in: *Proceedings, Part I, Web Information Systems Engineering – WISE 2020 – 21st International Conference*, Amsterdam, The Netherlands, October 20–24, 2020, Z. Huang, W. Beek, H. Wang, R. Zhou and Y. Zhang, eds, Lecture Notes in Computer Science, Vol. 12342, Springer, 2020, pp. 328–342. doi:10.1007/978-3-030-62005-9_24.

[79] M. Nayyeri, S. Vahdati, C. Aykul and J. Lehmann, 5* knowledge graph embeddings with projective transformations, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event*, February 2–9, 2021, AAAI Press, 2021, pp. 9064–9072, https://ojs.aaai.org/index.php/AAAI/article/view/17095.

[80] K. Noullet, R. Mix and M. Färber, KORE 50^{DYWC}: An evaluation data set for entity linking based on DBpedia, YAGO, Wikidata, and Crunchbase, in: *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020*, Marseille, France, May 11–16, 2020, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association, 2020, pp. 2389–2395, https://www.aclweb.org/anthology/2020.lrec-1.291/.

[81] I.L. Oliveira, R. Fileto, R. Speck, L.P. Garcia, D. Moussallem and J. Lehmann, Towards holistic Entity Linking: Survey and directions, *Information Systems* (2020), 101624.

[82] Oxford Online Dictionary, entity 2021, https://www.oxfordlearnersdictionaries.com/definition/american_english/entity.

[83] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical Report, Stanford InfoLab.

[84] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, *Semantic Web* **8**(3) (2017), 489–508. doi:10.3233/SW-160218.

[85] J. Pennington, R. Socher and C.D. Manning, Glove: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, a Meeting of SIGDAT, a Special Interest Group of the ACL*, Doha, Qatar, October 25–29, 2014, A. Moschitti, B. Pang and W. Daelemans, eds, ACL, 2014, pp. 1532–1543. doi:10.3115/v1/d14-1162.

[86] D. Perkins, Separating the Signal from the Noise: Predicting the Correct Entities in Named-Entity Linking, Master's thesis, Uppsala University, 2020.

[87] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, Deep contextualized word representations, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1, Long Papers*, New Orleans, Louisiana, USA, June 1–6, 2018, M.A. Walker, H. Ji and A. Stent, eds, Association for Computational Linguistics, 2018, pp. 2227–2237. doi:10.18653/v1/n18-1202.

[88] F. Petroni, A. Piktus, A. Fan, P.S.H. Lewis, M. Yazdani, N.D. Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel and S. Riedel, KILT: A benchmark for knowledge intensive language tasks, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online*, June 6–11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty and Y. Zhou, eds, Association for Computational Linguistics, 2021, pp. 2523–2544. doi:10.18653/v1/2021.naacl-main.200.

[89] V. Provatorova, S. Vakulenko, E. Kanoulas, K. Dercksen and J.M. van Hulst, *Named Entity Recognition and Linking on Historical Newspapers: UvA.ILPS & REL at CLEF HIPE 2020, Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum*, 2020.

[90] J. Raiman and O. Raiman, DeepType: Multilingual entity linking by neural type system evolution, in: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, New Orleans, Louisiana, USA, February 2–7, 2018, S.A. McIlraith and K.Q. Weinberger, eds, AAAI Press, 2018, pp. 5406–5413, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17148.

[91] L. Ratinov, D. Roth, D. Downey and M. Anderson, Local and global algorithms for disambiguation to Wikipedia, in: *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, Portland, Oregon, USA, 19–24 June, 2011, D. Lin, Y. Matsumoto and R. Mihalcea, eds, The Association for Computer Linguistics, 2011, pp. 1375–1384, https://www.aclweb.org/anthology/P11-1138/.

[92] S. Rijhwani, J. Xie, G. Neubig and J.G. Carbonell, Zero-shot neural transfer for cross-lingual entity linking, in: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, 2019, pp. 6924–6931. doi:10.1609/aaai.v33i01.33016924.

[93] D. Ringler and H. Paulheim, One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co, in: *KI 2017: Advances in Artificial Intelligence – 40th Annual German Conference on AI, Proceedings*, Dortmund, Germany, September 25–29, 2017, G. Kern-Isberner, J. Fürnkranz and M. Thimm, eds, Lecture Notes in Computer Science, Vol. 10505, Springer, 2017, pp. 366–372. doi:10.1007/978-3-319-67190-1_33.

[94] M. Röder, R. Usbeck, S. Hellmann, D. Gerber and A. Both, N^3 – a collection of datasets for named entity recognition and disambiguation in the NLP interchange format, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavik, Iceland, May 26–31, 2014, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani,

A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), 2014, pp. 3529–3533, http://www.lrec-conf.org/proceedings/lrec2014/summaries/856.html.

[95] M. Röder, R. Usbeck and A.N. Ngomo, GERBIL – benchmarking named entity recognition and linking consistently, *Semantic Web* **9**(5) (2018), 605–625. doi:10.3233/SW-170286.

[96] H. Rosales-Méndez, A. Hogan and B. Poblete, VoxEL: A benchmark dataset for multilingual entity linking, in: *Proceedings, Part II,the Semantic Web – ISWC 2018–17th International Semantic Web Conference*, Monterey, CA, USA, October 8–12, 2018, D. Vrandecic, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L. Kaffee and E. Simperl, eds, Lecture Notes in Computer Science, Vol. 11137, Springer, 2018, pp. 170–186. doi:10.1007/978-3-030-00668-6_11.

[97] H. Rosales-Méndez, A. Hogan and B. Poblete, Fine-Grained Entity Linking, *Journal of Web Semantics* (2020), 100600. doi:10.1016/j.websem.2020.100600.

[98] P. Rosso, D. Yang and P. Cudré-Mauroux, Beyond triplets: Hyper-relational knowledge graph embedding for link prediction, in: *WWW'20: The Web Conference 2020*, Taipei, Taiwan, April 20–24, 2020, Y. Huang, I. King, T. Liu and M. van Steen, eds, ACM/IW3C2, 2020, pp. 1885–1896. doi:10.1145/3366423.3380257.

[99] A. Sakor, I.O. Mulang, K. Singh, S. Shekarpour, M. Vidal, J. Lehmann and S. Auer, Old is gold: Linguistic driven approach for entity and relation linking of short text, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, June 2–7, 2019, S. Papers, J. Burstein, C. Doran and T. Solorio, eds, Association for Computational Linguistics, 2019, pp. 2336–2346. doi:10.18653/v1/n19-1243.

[100] A. Sakor, K. Singh, A. Patel and M. Vidal, Falcon 2.0: An entity and relation linking tool over Wikidata, in: *CIKM'20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event*, Ireland, October 19–23, 2020, M. d'Aquin, S. Dietze, C. Hauff, E. Curry and P. Cudré-Mauroux, eds, ACM, 2020, pp. 3141–3148. doi:10.1145/3340531.3412777.

[101] Ö. Sevgili, A. Shelmanov, M. Arkhipov, A. Panchenko and C. Biemann, Neural Entity Linking: A Survey of Models based on Deep Learning, 2020, CoRR, https://arxiv.org/abs/2006.00575 arXiv:2006.00575.

[102] W. Shen, J. Wang and J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, *IEEE Trans. Knowl. Data Eng.* **27**(2) (2015), 443–460. doi:10.1109/TKDE.2014.2327028.

[103] A. Singhal, Introducing the knowledge graph: Things, not strings, *Official Google blog* **5** (2012), 16.

[104] D. Sorokin and I. Gurevych, Context-aware representations for knowledge base relation extraction, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark, September 9–11, 2017, M. Palmer, R. Hwa and S. Riedel, eds, Association for Computational Linguistics, 2017, pp. 1784–1789. doi:10.18653/v1/d17-1188.

[105] D. Sorokin and I. Gurevych, Mixing context granularities for improved entity linking on question answering data across entity categories, in: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans*, Louisiana, USA, June 5–6, 2018, M. Nissim, J. Berant and A. Lenci, eds, Association for Computational Linguistics, 2018, pp. 65–75. doi:10.18653/v1/s18-2007.

[106] A. Sperduti and A. Starita, Supervised neural networks for the classification of structures, *IEEE Trans. Neural Networks* **8**(3) (1997), 714–735. doi:10.1109/72.572108.

[107] A. Spitz, J. Geiß and M. Gertz, So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks, in: *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, GeoRich@SIGMOD 2016*, San Francisco, California, USA, June 26–July 1, 2016, A. Züfle, B. Adams and D. Wu, eds, ACM, 2016, pp. 2:1–2:6. doi:10.1145/2948649.2948651.

[108] Y. Su, H. Sun, B.M. Sadler, M. Srivatsa, I. Gur, Z. Yan and X. Yan, On generating characteristic-rich question sets for QA evaluation, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, Austin, Texas, USA, November 1–4, 2016, J. Su, X. Carreras and K. Duh, eds, The Association for Computational Linguistics, 2016, pp. 562–572. doi:10.18653/v1/d16-1054.

[109] T.P. Tanon, G. Weikum and F.M. Suchanek, YAGO 4: A reason-able knowledge base, in: *The Semantic Web – 17th International Conference, ESWC 2020, Proceedings*, Heraklion, Crete, Greece, May 31–June 4, 2020, A. Harth, S. Kirrane, A.N. Ngomo, H. Paulheim, A. Rula, A.L. Gentile, P. Haase and M. Cochez, eds, Lecture Notes in Computer Science, Vol. 12123, Springer, 2020, pp. 583–596. doi:10.1007/978-3-030-49461-2_34.

[110] K. Teru, E. Denis and W. Hamilton, Inductive relation prediction by subgraph reasoning, in: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event*, 13–18 July 2020, Proceedings of Machine Learning Research, Vol. 119, PMLR, 2020, pp. 9448–9457, http://proceedings.mlr.press/v119/teru20a.html.

[111] A. Thawani, M. Hu, E. Hu, H. Zafar, N.T. Divvala, A. Singh, E. Qasemi, P.A. Szekely and J. Pujara, Entity linking to knowledge graphs to infer column types and properties, in: *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching Co-Located with the 18th International Semantic Web Conference, SemTab@ISWC 2019*, Auckland, New Zealand, October 30, 2019, E. Jiménez-Ruiz, O. Hassanzadeh, K. Srinivas, V. Efthymiou and J. Chen, eds, CEUR Workshop Proceedings, Vol. 2553, CEUR-WS.org 2019, pp. 25–32, http://ceur-ws.org/Vol-2553/paper4.pdf.

[112] C. Tsai and D. Roth, Cross-lingual wikification using multilingual embeddings, in: *NAACL HLT 2016, the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego California, USA, June 12–17, 2016, K. Knight, A. Nenkova and O. Rambow, eds, The Association for Computational Linguistics, 2016, pp. 589–598. doi:10.18653/v1/n16-1072.

[113] R. Usbeck, A.N. Ngomo, B. Haarmann, A. Krithara, M. Röder and G. Napolitano, 7th open challenge on Question Answering over Linked Data (QALD-7), in: *Semantic Web Challenges – 4th SemWebEval Challenge at ESWC 2017, Revised Selected Papers*, Portoroz, Slovenia, May 28–June 1, 2017, M. Dragoni, M. Solanki and E. Blomqvist, eds, Communications in Computer and Information Science, Vol. 769, Springer, 2017, pp. 59–69. doi:10.1007/978-3-319-69146-6_6.

[114] R. Usbeck, A.N. Ngomo, M. Röder, D. Gerber, S.A. Coelho, S. Auer and A. Both, AGDISTIS – agnostic disambiguation of named entities using linked open data, in: *ECAI 2014–21st European Conference on Artificial Intelligence – Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, Prague, Czech Republic, 18–22 August, 2014, T. Schaub, G. Friedrich and B. O'Sullivan, eds, Frontiers in Artificial Intelligence and Applications, Vol. 263, IOS Press, 2014, pp. 1113–1114. doi:10.3233/978-1-61499-419-0-1113.

[115] J.M. van Hulst, F. Hasibi, K. Dercksen, K. Balog and A.P. de Vries, REL: An entity linker standing on the shoulders of giants, in: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event*, China, July 25–30, 2020, J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen and Y. Liu, eds, ACM, 2020, pp. 2197–2200. doi:10.1145/3397271.3401416.

[116] T. van Veen, J. Lonij and W.J. Faber, Linking named entities in Dutch historical newspapers, in: *Metadata and Semantics Research – 10th International Conference, MTSR 2016, Proceedings*, Göttingen, Germany, November 22–25, 2016, E. Garoufallou, I.S. Coll, A. Stellato and J. Greenberg, eds, Communications in Computer and Information Science, Vol. 672, 2016, pp. 205–210. doi:10.1007/978-3-319-49157-8_18.

[117] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA, December 4–9, 2017, I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan and R. Garnett, eds, 2017, pp. 5998–6008, http://papers.nips.cc/paper/7181-attention-is-all-you-need.

[118] O. Vinyals, M. Fortunato and N. Jaitly, Pointer networks, in: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, Montreal, Quebec, Canada, December 7–12, 2015 C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett, eds, 2015, pp. 2692–2700, http://papers.nips.cc/paper/5866-pointer-networks.

[119] D. Vrandecic, Building a multilingual Wikipedia, *Commun. ACM* **64**(4) (2021), 38–41. doi:10.1145/3425778.

[120] D. Vrandečić and M. Krötzsch, Wikidata: A free collaborative knowledgebase, *Commun. ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.

[121] W3C Microposts Community Group, *Making Sense of Microposts (#Microposts2016)*, 2016, http://microposts2016.seas.upenn.edu/challenge.html.

[122] P. Wang, J. Han, C. Li and R. Pan, Logic attention based neighborhood aggregation for inductive knowledge graph embedding, in: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, 2019, pp. 7152–7159. doi:10.1609/aaai.v33i01.33017152.

[123] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li and J. Tang, KEPLER: A unified model for knowledge embedding and pre-trained language representation, *Trans. Assoc. Comput. Linguistics* **9** (2021), 176–194, https://transacl.org/ojs/index.php/tacl/article/view/2447. doi:10.1162/tacl_a_00360.

[124] G. Weikum, X.L. Dong, S. Razniewski and F.M. Suchanek, Machine knowledge: Creation and curation of comprehensive knowledge bases, *Found. Trends Databases* **10**(2–4) (2021), 108–490. doi:10.1561/1900000064.

[125] Wikimedia Foundation, Index of /wikidatawiki/entities/, 2020-08-21. https://dumps.wikimedia.org/wikidatawiki/entities/.

[126] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne et al., The FAIR guiding principles for scientific data management and stewardship, *Scientific data* **3**(1) (2016), 1–9.

[127] L. Wu, F. Petroni, M. Josifoski, S. Riedel and L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, November 16–20, 2020, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 6397–6407. doi:10.18653/v1/2020.emnlp-main.519.

[128] T. Wu, A. Khan, H. Gao and C. Li, Efficiently Embedding Dynamic Knowledge Graphs, 2019, CoRR, http://arxiv.org/abs/1910.06708 arXiv:1910.06708.

[129] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and S.Y. Philip, A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[130] X. Yang, X. Gu, S. Lin, S. Tang, Y. Zhuang, F. Wu, Z. Chen, G. Hu and X. Ren, Learning dynamic context augmentation for global entity linking, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, Hong Kong, China, November 3–7, 2019, K. Inui, J. Jiang, V. Ng and X. Wan, eds, Association for Computational Linguistics, 2019, pp. 271–281. doi:10.18653/v1/D19-1026.

[131] Z. Yang, Z. Dai, Y. Yang, J.G. Carbonell, R. Salakhutdinov and Q.V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, Vancouver, BC, Canada, 8–14 December, 2019, H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox and R. Garnett, eds, 2019, pp. 5754–5764, http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.

[132] W. Yih, M. Richardson, C. Meek, M. Chang and J. Suh, The value of semantic parse labeling for knowledge base question answering, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Short Papers*, Berlin, Germany, August 7–12, 2016, Vol. 2, The Association for Computer Linguistics, 2016. doi:10.18653/v1/p16-2033.

[133] F. Yu and V. Koltun, Multi-scale context aggregation by dilated convolutions, in: *4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, eds, San Juan and Rico, Puerto, May 2–4, 2016 2016, http://arxiv.org/abs/1511.07122.

[134] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, Quality assessment for linked data: A survey, *Semantic Web* **7**(1) (2016), 63–93. doi:10.3233/SW-150175.

[135] Z. Zhang, J. Cai, Y. Zhang and J. Wang, Learning hierarchy-aware knowledge graph embeddings for link prediction, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7–12, 2020, AAAI Press, 2020, pp. 3065–3072, https://aaai.org/ojs/index.php/AAAI/article/view/5701.

[136] S. Zhou, S. Rijhwani, J. Wieting, J.G. Carbonell and G. Neubig, Improving candidate generation for low-resource cross-lingual entity linking, *Trans. Assoc. Comput. Linguistics* **8** (2020), 109–124, https://transacl.org/ojs/index.php/tacl/article/view/1906. doi:10.1162/tacl_a_00303.

[137] S. Zwicklbauer, C. Seifert and M. Granitzer, DoSeR – a knowledge-base-agnostic framework for entity disambiguation using semantic embeddings, in: *The Semantic Web. Latest Advances and New Domains – 13th International Conference, ESWC 2016, Proceedings*, Heraklion, Crete, Greece, May 29–June 2, 2016, H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S.P. Ponzetto and C. Lange, eds, Lecture Notes in Computer Science, Vol. 9678, Springer, 2016, pp. 182–198. doi:10.1007/978-3-319-34129-3_12.