Published: 25 February 2026
Online AM: 31 January 2026
Accepted: 22 January 2026
Revised: 20 December 2025
Received: 09 February 2024

Citation in BibTeX format

SURVEY

Entity Linking with Wikidata: A Systematic Literature Review

PHILIPP SCHARPF, The University of Göttingen, Gottingen, Niedersachsen, Germany

I review, develop, and evaluate Mathematical Entity Linking (MathEL) methods and applications. Methods include STEM document (formula & identifier) annotation recommendation, formula concept retrieval (classification & clustering), and nearest-neighbor retrieval. Applications include mathematical (STEM) document classification & clustering, Wikipage readability & accessibility, formula search, question answering, and question generation. Particularly, I am focused on grounding the formula semantics to the Wikidata knowledge graph.

CORINNA BREITINGER, The University of Göttingen, Gottingen, Niedersachsen, Germany

ANDREAS SPITZ, University of Konstanz, Konstanz, Baden-Wurttemberg, Germany

NORMAN MEUSCHKE, The University of Göttingen, Gottingen, Niedersachsen, Germany

ANDRÉ GREINER-PETTER, The University of Göttingen, Gottingen, Niedersachsen, Germany

MORITZ SCHUBOTZ, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Baden-Wurttemberg, Germany

View all

Entity Linking with Wikidata: A Systematic Literature Review

PHILIPP SCHARPF, Computer Science, University of Göttingen, Göttingen, Germany
CORINNA BREITINGER, Computer Science, University of Göttingen, Göttingen, Germany
ANDREAS SPITZ, University of Konstanz, Konstanz, Germany
NORMAN MEUSCHKE, Computer Science, University of Göttingen, Göttingen, Germany
ANDRÉ GREINER-PETTER, University of Göttingen, Göttingen, Germany
MORITZ SCHUBOTZ, FIZ Karlsruhe Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
BELA GIPP, Computer Science, University of Göttingen, Göttingen, Germany

This article provides a comprehensive systematic review of the literature on entity linking using Wikidata as the grounding knowledge base. Our review extends the scope of previous studies from two to eight dimensions of entity linking, which we classify into the following categories: definitions, tasks, types, domains, approaches, datasets, applications, and challenges. We find that datasets primarily address question-answering and news domains but underutilize Wikidata's capabilities for hyper-relations, multilingualism, and time dependence. The research gaps we identify include the need for more robust datasets, hybrid methods combining rule-based and learning-based approaches, and improved handling of ambiguity, sparse entity types, data noise, and knowledge graph evolution.

CCS Concepts: • **Information systems → Information retrieval;**

Additional Key Words and Phrases: Entity linking, named entity recognition, wikidata

ACM Reference Format:
Philipp Scharpf, Corinna Breitinger, Andreas Spitz, Norman Meuschke, André Greiner-Petter, Moritz Schubotz, and Bela Gipp. 2026. Entity Linking with Wikidata: A Systematic Literature Review. *ACM Comput. Surv.* 58, 9, Article 227 (February 2026), 50 pages. https://doi.org/10.1145/3795134

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – grants 437179652 and 554559555, as well as the Lower Saxony Ministry of Science and Culture.
Authors' Contact Information: Philipp Scharpf, Computer Science, University of Göttingen, Göttingen, Niedersachsen, Germany; e-mail: scharpf@gipplab.org; Corinna Breitinger (corresponding author), Computer Science, University of Göttingen, Göttingen, Niedersachsen, Germany; e-mail: corinna.breitinger@uni-goettingen.de; Andreas Spitz, University of Konstanz, Konstanz, Baden-Württemberg, Germany; e-mail: andreas.spitz@uni-konstanz.de; Norman Meuschke, Computer Science, University of Göttingen, Göttingen, Niedersachsen, Germany; e-mail: meuschke@uni-goettingen.de; André Greiner-Petter, University of Göttingen, Göttingen, Niedersachsen, Germany; e-mail: greinerpetter@gipplab.org; Moritz Schubotz, FIZ Karlsruhe Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Baden-Württemberg, Germany; e-mail: moritz.schubotz@fiz-karlsruhe.de; Bela Gipp, Computer Science, University of Göttingen, Göttingen, Niedersachsen, Germany; e-mail: gipp@uni-goettingen.de.

ACM 0360-0300/2026/02-ART227
https://doi.org/10.1145/3795134

1 Introduction

Named entities are terms that represent real-world objects, such as people, places, organizations, and dates, which can be identified and classified in text [53]. Identifying named entities is a crucial component of natural language processing and information retrieval systems for tasks such as question answering, semantic search, and sentiment analysis [80]. Named entity recognition or entity linking refers to associating and disambiguating mentions of entities of primary entity types, such as persons, organizations, locations, dates, and times, with representations in a knowledge base or knowledge graph [22]. Entity linking is likewise integral to information extraction, i.e., obtaining structured information or queries from unstructured texts such as newspaper articles or encyclopedias[53].

Entity linking is challenging due to the following:

— Synonymy: Multiple terms (entity "surface forms") can refer to the same entity concept;
— Polysemy: An entity name can refer to multiple concepts [48];
— Missing entities: There are surface forms without a corresponding entity in the target knowledge base, commonly labeled as "Not In Lexicon" [14].

These challenges require entity linking to exploit the entity context or additional information to disambiguate and create new knowledge base items if necessary. Unlike deduplication, which merges duplicates within a dataset, or record linkage, which matches entities across datasets, entity linking resolves ambiguities and enriches text with structured context, enhancing information retrieval and knowledge management [72].

Current reviews on entity linking or named entity recognition [69, 72] focus on Wikipedia as the grounding knowledge graph. However, Wikipedia as a knowledge graph is merely semi-structured and language-dependent. Wikidata was introduced in 2012 to address the shortcomings of Wikipedia by providing structured and multilingual representations for concept entities [84]. New entity linking approaches have gradually taken advantage of these new capabilities by also including Wikidata as a grounding knowledge base. However, so far, only one literature review addressed entity linking with Wikidata [50] by surveying approaches and datasets, but omitting dimensions that other reviews on entity linking have investigated, such as entity linking definitions [41, 69, 72], tasks [86], types [53], domains [53], applications [69, 72], and challenges [41, 72].

This article discusses the definitions, tasks, approaches, datasets, entity types, primary application domains, promising practical applications, and open challenges of entity linking. We systematically derive an entity linking framework and identify the research gaps in each "review dimension" (see Figure 2).[1] Our research updates and extends the previous review [50] of entity linking with Wikidata from 2 to 8 review dimensions. This is essential for researchers to gain a thorough understanding of the field and reflects the same review dimensions that were used by other reviews in which Wikipedia, instead of Wikidata, was the grounding knowledge base.

1.1 Overview

This section *introduces* Wikidata, defines our research objectives and research questions, and derives our research tasks. The remainder of this article is structured as follows. Section 2 discusses related reviews and our search *methodology*. Section 3 presents the *results* for selected review dimensions that meet our research goals, answer our research questions, and complement the previous literature review (see Section 1.5 for an explanation of the individual dimensions). The section also includes a *discussion* through which we derive a generalized framework, highlight or critique research

[1]We refer to definitions, tasks, approaches, datasets, applications, and challenges as "review dimensions".

methods, and identify gaps or disagreements in the existing research literature. Section 4 *concludes* our analysis and presents our plans for *future work*.

The terms named entity recognition, named entity recognition and classification, named entity disambiguation, named entity linking, entity disambiguation, and entity linking are often used interchangeably. In this article, we summarily refer to these tasks as entity linking.

1.2 Entity Linking

Entity linking or named entity linking is the natural language processing task of mapping entities in texts to an entry in a knowledge base or node in a knowledge graph [25]. Thereby, entity linking transforms unstructured ambiguous text (noun phrases, questions, etc.) to structured sets of semantic entity references [66]. Entity reference sets are structured lists of correspondences between entity mention names in the text and a "semantic" representation of their meaning in the knowledge graph. Entity linking with Wikipedia or Wikidata is also called "Wikification" [47]. The entity linking task is important for many information retrieval and natural language processing applications, such as semantic search, question answering, chatbots, and knowledge base Population [80]. One of the earliest and widely adopted definitions developed at the **sixth Message Understanding Conference (MUC-6)**,[2] limited named entity recognition to persons, locations, organizations, temporal expressions (dates, times), and quantities, such as monetary values or percentages [23]. Later definitions extended the classes to products, financial entities, films, scientists, and so on. [62].

Entity linking approaches can be broadly categorized into rule-based (including dictionary-based), statistical, and neural methods [69]. Rule-based methods rely on hand-crafted rules for detecting and disambiguating entity mentions. In contrast, statistical methods build models based on features derived from training data. Neural methods represent entity mentions and knowledge base entities in a shared vector space, and use these representations for detecting and disambiguating entity mentions [36]. While classical named entity recognition or entity linking techniques have been researched for decades, the integration of Deep Neural Networks has led to significant advancements in the past decade [86].

1.3 Wikidata

Wikidata is a free, collaborative multilingual knowledge base that supports factual information in Wikipedia. Wikidata was launched in 2012 to complement Wikipedia with structured secondary information and language-independent representations for knowledge concepts [84]. Wikidata gradually replaced an earlier project, Freebase, as the open core of the Google knowledge graph. Google originally intended Freebase to serve as a: "Wikipedia for structured data" [81]. As Wikidata evolved, its popularity and active community convinced Google to shift from Freebase to Wikidata [81].

The data model of Wikidata consists of items that contain statements, which, in turn, contain claims with qualifiers and references. For example, the item "Eiffel Tower" (query ID Q243) [3] includes the claim "height" (property ID P2048) with the property value 300 meters and a qualifier "architectural height" (Q24192182) supported by several URL references. Having structured information allows for structured queries, e.g., using the SPARQL query language.[4]

Wikidata's active community makes it a highly relevant knowledge base for entity linking. The increase in metrics, such as the number of active editors (cf. Figure 1(a)) and the frequency of item edits (cf. Figure 1(b)) attest to Wikidata's active community.

[2] https://cs.nyu.edu/grishman/muc6.html
[3] https://www.wikidata.org/wiki/Q243
[4] https://query.wikidata.org/

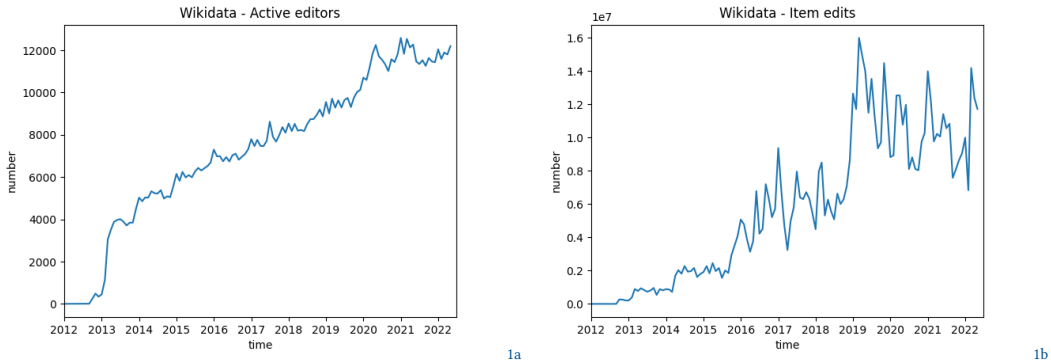


Fig. 1. Wikidata editor and item statistics: Active editors (1(a)) and item edits (1(b)).

1.4 Entity Linking Using Wikidata

1.4.1 Challenges. Entity linking using Wikidata is distinguished from entity linking using other knowledge bases by several unique and challenging properties, which we describe in the following.

Dynamic Nature. Wikidata's rapid and ongoing updates by a global community make it a dynamic knowledge base. This evolving nature poses challenges for keeping entity linking systems current and dealing with inconsistencies [12]. For example, the Wikidata property "date of birth" has changed over time by shifting from loosely formatted year-only values to more precise date entries with explicit calendar models.

Multilinguality. Wikidata supports numerous languages with entity labels and descriptions, enabling entity linking systems to handle complex disambiguation tasks on multilingual datasets [32]. For example, the German alias "Amerika" for the "United States" can be misleading in English, where "America" may refer to the continent rather than the country, causing ambiguity for entity linking systems.

Alias Diversity. Entities in Wikidata often have many aliases, including culturally specific names, which requires sophisticated modeling of name variations [52]. For example, a system may fail to link a query for "The King of Pop" to 'Michael Jackson" if that culturally specific alias is not part of its reference set.

Hierarchical Structure. The intricate relationships and hierarchies in Wikidata demand advanced graph-based techniques to utilize its semantic richness [66]. For example, a scholarly article on 'Homo sapiens" could be mislinked to the general concept of humans rather than the specific taxonomic entity if hierarchical context is not modeled well.

Breadth and Depth. Covering a wide array of domains, including niche and specialized topics, Wikidata introduces challenges in disambiguating entities with overlapping or sparse contextual clues [50]. For example, a historical biography mentioning 'Duke of York" could be incorrectly linked to the current titleholder instead of a historical figure from the 15th century.

These factors collectively make entity linking using Wikidata a sophisticated task, requiring robust solutions that leverage its multilingual, hierarchical, and evolving nature while mitigating alias variability and coverage breadth.

1.4.2 Methods. The challenges of entity linking using Wikidata are tackled by advanced methods. In the following, we describe the most characteristic problem categories and solution approaches.

Joint Entity and Relation Linking. Unlike traditional entity linking systems that resolve mentions to human-readable Wikipedia page titles, Wikidata entity linking targets machine-readable entity item QIDs (e.g., Q76 for Barack Obama) and edge property PIDs (e.g., P27 for country of citizenship). This requires candidate generation pipelines that match surface forms to opaque identifiers and respect Wikidata's schema. Falcon 2.0 [66] jointly links entity and relation mentions by resolving to QIDs and PIDs, leveraging the SPARQL protocol and RDF query language on Wikidata's triple structure, e.g.,

```
SELECT ?object WHERE {wd:Q7266513 wdt:P3362 ?object.}
```

to find the `</entity/Q7266513></prop/direct/P3362>?object` for the subject item Q7266513 and the predicate property P3362.

KBPearl [40] builds a weighted semantic graph where nodes represent noun phrases, relation phrases, and their corresponding candidate entities or predicates. Edges between nodes are assigned similarity scores, which denote the likelihood that a noun phrase refers to an entity and that a relation phrase maps to a predicate, based on alias overlaps and keyphrase similarity. The system formulates the disambiguation task as a dense subgraph problem, aiming to extract a subgraph that connects each noun phrase and relation phrase to exactly one candidate entity or predicate. To capture global coherence, the density of the subgraph is defined as the minimum weighted degree of the set of nodes in the subgraph over the sum of weights of edges incident to a node. A greedy pruning algorithm is used to remove low-coherence candidates while maintaining these linking constraints, resulting in a coherent and contextually consistent mapping from text to knowledge base entries.

Modeling Property Graphs. Wikidata's knowledge representation is inherently graph-based and modeled as a labeled directed graph, where nodes represent entities, classes, and literals, and directed edges represent properties connecting head and tail nodes. These properties encode rich semantic and alias relationships but are noisy and inconsistent due to the collaborative nature of Wikidata.

Mulang et al. [52] propose an entity linking approach that integrates this property graph context into a neural architecture by leveraging a local subgraph, which is centered around a candidate entity. For each surface form, candidate entities are retrieved via semantic search over entity labels and aliases. The disambiguation is modeled as a classification function that depends on alias-based contextual information derived from the subgraph using attention-based BiLSTM networks. This modeling directly incorporates graph structure into the neural decision process, improving disambiguation for non-standard and long-tail entities.

VCG [75] further introduces a variable-context mechanism that selectively invokes property graph embeddings when they are expected to improve disambiguation, balancing accuracy with computational efficiency.

Learning in Knowledge Graphs. Entity linking over Wikidata increasingly benefits from representation learning methods [5] that encode entities and relations of the knowledge graph as continuous vector representations. These embeddings aim at capturing both the semantic context–such as the lexical similarity between entity labels and aliases–and the structural context, including an entity's position within the graph and its connectivity through relation types. In practice, this

[5]Note that we exclude graph representation learning in this review, since it is a very broad topic that deserves a separate literature review.

allows learning-based models to recognize, for instance, that two entities with different labels but similar neighborhoods (e.g., both subclass of the same parent or connected to the same properties) should be embedded nearby in vector space.

Boros et al. [5] incorporate such graph-derived representations into a neural reranking stage, where candidate entities initially retrieved via surface similarity are re-evaluated using a deep neural network that takes both text and entity embeddings as input. This architecture demonstrates improved robustness to OCR-induced errors and morphological complexity, particularly in multilingual and historical document settings.

Similarly, Hamdi et al. [26] fine-tune multilingual BERT encoders that implicitly exploit structural features of Wikidata—such as entity frequency and alias distributions—by integrating them into the fine-tuning dataset. These approaches reflect a broader shift toward heterogeneous graph representation learning, where structurally diverse features–such as alias relations, temporal qualifiers, and class hierarchies–are encoded jointly. This enables more accurate disambiguation in real-world conditions characterized by noise, sparsity, and low-resource languages.

Handling Multilingual Labels. Wikidata supports labels and aliases in hundreds of languages, necessitating entity linking methods that normalize multilingual mentions to the same QID. One core difficulty lies in aligning noisy or morphologically variable surface forms across languages to a unified identifier.

Hamdi et al. [26] address this by curating the NewsEye dataset, which includes Wikidata QID annotations for named entities across four European languages. Their evaluation focuses on cross-lingual mention resolution under OCR degradation and spelling variance, combining language-specific mention-entity probability tables with BERT encoders.

Boros et al. [5] further investigate multilingual EL under historical OCR noise, stacking Transformer layers on top of pretrained BERT models and projecting mention spans into a shared space with MUSE multilingual word embeddings. They highlight that language morphology and subword tokenization errors, such as deletions and character-level corruption, strongly affect EL performance under noisy conditions.

1.5 Research Objectives

Our *Research Goal* is to provide a comprehensive overview of entity linking based on the English subset of Wikidata by systematically reviewing and discussing entity linking definitions, tasks, types, domains, approaches, datasets, applications, and challenges (eight review dimensions).

Our analysis is motivated by the following *Research Questions*:

(1) What do researchers need to know about entity linking with Wikidata in terms of definitions, tasks, types, domains, approaches, datasets, and applications?
(2) Where is the research need and potential (research gaps)?

We answer these Research Questions in Section 3.4.

Overall, we seek to identify the prevailing entity linking definitions and tasks in the current literature to distill a general entity linking pipeline scheme. This framework will incorporate insights and methodologies from the reviewed literature to reveal a systematic view on entity linking. Moreover, our objective is to outline the research gaps pertaining to current entity linking types, domains, approaches, datasets, applications, and challenges to guide future research on entity linking. Therefore, we derive the following ***Research Tasks*** **(RTs)**

— **RT 1. Entity linking review**: Perform a systematic literature search, report, and discuss review results in each review dimension.
— **RT 2. Entity linking framework**: Distil the retrieved entity linking definitions and tasks to derive a theoretical framework for entity linking with Wikidata.

Fig. 2. Overview of the 8 included entity linking review dimensions with examples.

Table 1. Research Tasks (RTs) and the Review Dimensions They Include

RT / Dimension	Def.	Tasks	Types	Dom.	Appr.	Datas.	Appl.	Chall.
1 Review	x	x	x	x	x	x	x	x
2 Framework	x	x	-	-	-	-	-	-
3 Gaps	-	-	x	x	x	x	x	x

— **RT 3. Entity linking research gaps**: Identify research gaps in types, domains, approaches, datasets, applications, and challenges to guide future research.

Table 1 illustrates which individual review dimensions (definitions, tasks, types, domains, approaches, datasets, applications, challenges) serve to complete our research tasks (review, framework, gaps). Note that the goal of RT 2 is consolidating entity linking definitions into a framework and exploring specifically how Wikidata is used in this framework. Thus, the framework is not generally about entity linking but focusses on Wikidata and how entity linking is defined specifically in the analyzed Wikidata entity linking articles. Thus, the framework is Wikidata-specific and aims to support a unified understanding of its role in entity linking research.

Figure 2 summarizes the review dimensions using examples.

The review dimensions are characterized as follows:

— *Definitions*. We gather different descriptions of the entity linking process, such as 'detecting mentions of entities from a knowledge base in free text.'

— *Tasks*. We identify tasks and subtasks of the entity linking process, such as candidate generation and ranking.
— *Approaches*. We discuss the most popular approaches to entity linking with Wikidata, e.g., 'Falcon' for joint entity and relation linking.
— *Datasets*. We introduce the most frequently used datasets for entity linking with Wikidata, e.g., 'HIPE' consisting of annotated historical newspapers.
— *Types*. We determine the most important entity linking entity types, such as person, organization, and location and more.
— *Domains*. We highlight the prevailing entity linking domains, such as news or medical texts.
— *Applications*. We explain the most researched applications, e.g., question answering.
— *Challenges*. We elaborate on the greatest challenges for entity linking with Wikidata, e.g., knowledge graph evolution.

2 Methodology

In this section, we outline our review methodology. We follow the guidelines of [35], which is a condensation of the well-known review procedure guidelines by Kitchenham and Charters [33]. We start by discussing related reviews, then detail our systematic search and selection process, and conclude with some statistics on the publications included in our review.

2.1 Related Reviews

Most surveys on entity linking or named entity recognition discuss approaches that use DBpedia, Wikipedia, Freebase, or YAGO as grounding knowledge graphs; some are knowledge graph agnostic. Only one recent review specifically focuses on entity linking with Wikidata. We will place our review in relation to these surveys at the end of this section.

Classical Entity Linking. In 2007, Nadeau and Sekine published a comprehensive review of fifteen years of research (from 1991 to 2006) in the field of named entity recognition and classification.[6] The report explores entity types, domains (or genres) and contrasts between rules-based and Machine Learning techniques. Additionally, it outlines entity features and encodings, spanning word-level, dictionary-level, and corpus-level representations, and touches upon evaluation methodologies [53]. In 2015, Shen et al. provided a review that primarily covers the definitions, techniques, challenges, and solutions in classical entity linking. Additionally, the authors provide an overview and analysis of the main entity linking approaches, discuss various applications, evaluate entity linking systems, and suggest future directions [72]. In the same year, Ling et al. published a review that identifies design challenges for entity linking and proposes promising techniques to address them, including deep neural networks and joint inference. They developed a simple, modular, unsupervised entity linking system and compared it to two state-of-the-art systems on 9 datasets. Additionally, the authors provide a detailed description of key entity linking tasks, such as mention extraction, candidate generation, entity type prediction, entity coreference, and coherence [41]. In 2018, Goyal et al. published another systematic review on the evolution of named entity recognition and classification research [22].

Neural Entity Linking. Yadav et al. published a survey on recent advances in entity recognition through deep learning models in 2018. The authors compare neural network architectures for entity recognition to previous supervised or semi-supervised machine learning algorithms. They highlight the improvements achieved by neural networks and show how applying lessons learned from earlier

[6]Recall that we subsume named entity recognition and classification under entity linking.

Table 2. Comparison of Literature Reviews on Entity Linking with English Wikidata

	Möller et al. [50]	**This review**
Definitions	-	x
Tasks	-	x
Types	-	x
Domains	-	x
Approaches	x	x
Datasets	x	x
Applications	-	x
Challenges	-	x

feature-based entity recognition systems can drive even greater progress [86]. In 2020, Al-Moslmi et al. presented a literature overview focussing on named entity extraction for knowledge graph population. The report covers the recent achievements of entity recognition, named entity detection, and named entity linking approaches. The authors identify the need for standard techniques to assess, evaluate, and compare entity recognition methods. They also emphasize the transition in entity linking development from stepwise pipelines to end-to-end architectures that consider context at each stage, with deep learning models being particularly successful [1]. In 2022, Sevgili et al. conducted a comprehensive review of neural deep learning for entity linking. The survey covers neural approaches from 2015 to 2020, and compares 30 different techniques on 9 diverse datasets, summarizing design features (e.g., candidate generation and ranking) and embedding approaches. The authors further compare the performances of classical and neural entity linking on established benchmarks and classify the components into common themes, such as joint entity recognition and linking, global linking, and domain-independent methods, including zero-shot and distant supervision methods and cross-lingual techniques. Finally, the authors discuss various entity linking applications and use cases [69].

Review Gap. None of the aforementioned reviews focuses on entity linking with Wikidata as the grounding knowledge base. However, a recent survey by Möller et al. [50] discusses 16 approaches and 11 datasets specifically built for Wikidata. We extend this literature review by a systematic examination of additional review dimensions, which we identified in the other reviews: Entity linking definitions [41, 69, 72], tasks [86], types [53], domains [53], applications [69, 72], and challenges [41, 72]. Table 2 shows that the literature review of Möller et al. [50] only covered 1/4 of the dimensions that we include in our review. Our motivation for addressing these previously ignored dimensions is the following:

(1) Other reviews of entity linking with Wikipedia (discussed above) include these dimensions; we consider them essential for a comprehensive overview of the field;
(2) We would like to guide future research on entity linking by providing Wikidata knowledge base-specific definitions and tasks, describing domains & types, and presenting entity linking applications and challenges (research gaps that need to be tackled).

Note that only 2 of the 8 dimensions (definitions and tasks) are generalizable or agnostic to the knowledge base. The remaining 6 (domains and types, approaches and datasets, applications and challenges) largely depend on the knowledge base used, that is, Wikidata in our case.

Table 3. Number of Search Results for the Query ‘ “Entity Linking” AND Wikidata’

Input / Source	Web of Science	ACM DL	ACL Anth.	IEEE Xplore	Springer Link	Science Direct	dblp	Total
“entity linking” AND wikidata	6	132	34	1	298	57	10	538

2.2 Publication Search and Selection

Our publication search and selection process comprised the following steps:

— Eligible papers were required to be available through one of the following academic search engines or repositories: *Web of Science, ACM DL, ACL Anthology, IEEE Xplore, Springer Link, Science Direct*, or *dblp*. These sources were the most frequently used in prior reviews, thereby ensuring a degree of consistency and quality. Consequently, non–peer-reviewed preprint servers, such as arXiv, were excluded.
— We used the same query as Möller et al. in their review of entity linking: ‘ “Entity Linking” AND Wikidata’ restricted to publication abstracts.

All retrieved results (see Table 3) were examined and manually filtered according to our predefined exclusion criteria. A publication was excluded if it met at least one of the following conditions:

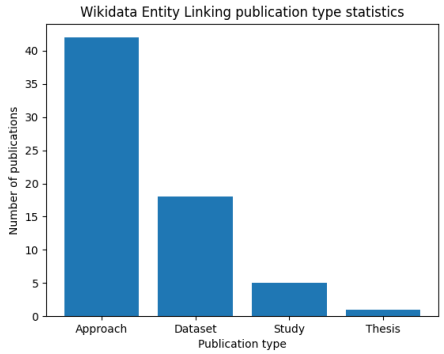
— it was not written in English;
— it did not primarily use the English Wikidata, as publications focusing on multilingual Wikidata applications were excluded as a distinct research niche;
— it addressed the knowledge graph population in general, a broad topic warranting a separate review;
— it did not mention “Entity Linking” and Wikidata’ in the abstract, did not present experiments, or did not discuss entity linking with Wikidata. Many publications in this category referred to Wikidata only briefly, for example, as a potential knowledge base for entity linking.

Conversely, publications were included if they were written in English and either presented experiments using Wikidata as a knowledge base for entity linking tasks or explicitly discussed entity linking with Wikidata.

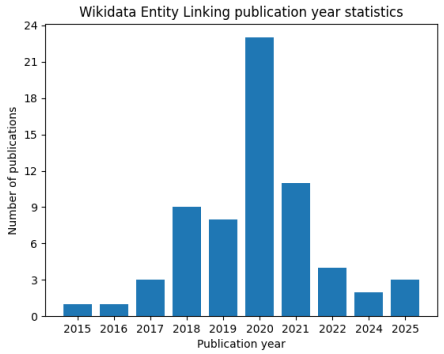
Finally, to update our review with the most recent publications, we performed forward citation snowballing. In November 2025, we conducted forward snowballing on the 54 publications we initially included by using Google Scholar’s “cited by” statistics to find more recent papers. The 54 publications had been cited by 2,383 papers, of which 1,193 mentioned the keywords Entity Linking and Wikidata in the abstract or full text. After removing duplicates and applying the same search and inclusion/exclusion criteria detailed above, we identified 11 additional publications meeting all criteria. Lastly, we forward snowballed these newly identified 11 publications, which had received 66 citations; however, no additional relevant publications were identified among these citations. Thus, **65 publications** are included in our review. For the full table tracking our snowballing procedure, see https://zenodo.org/records/17839768.

2.3 Publication Types

Figure 3(a) shows the distribution of publication types in the collection we review. We distinguish four publication categories: Approach, dataset, study, and thesis. Approach means that the publication describes one or more methods. Dataset denotes that it primarily introduces a new



3a



3b

Fig. 3. Publication statistics of this review. Figure 3(a) shows the types of the reviewed publications, while Figure 3(b) shows the distribution of the publication year.

Table 4. Overview of Entity Linking Review Statistics for Types, Domains, Applications, and Challenges with Publication Count

Types	**Domains**	**Applications**	**Challenges**
Person: 38	News: 25	Knowledge Graph Population: 11	Specific Approaches: 18
Organization: 37	Articles: 10	Knowledge Graph Question Answering: 9	Knowledge Base or Knowledge Graph evolution: 17
Location: 36	Questions: 8	Relation Extraction: 6	Quality of Datasets: 14
Product: 15	Medicical Texts: 3	Entity Extraction: 1	Entity Ambiguity: 12
Time: 12	Tweets: 2	Stance Detection: 1	Data Sparsity and Noise: 18

dataset. Study refers to metastudies that discuss approaches or datasets from other researchers. Most publications present approaches (methods) or datasets.

3 Results and Discussion

In this section, we first present, describe, and analyze the review results [7] across our chosen dimensions (RT 1 - Results). Next, we introduce a framework for the entity linking pipeline, encompassing relevant definitions and tasks frequently found in the literature (RT 2 - Framework). Finally, we identify research gaps in each review dimension, compare them with prior reviews on Wikidata entity linking, and synthesize our insights into research guidelines (RT 3 - Research Gaps).

3.1 RT 1. Entity Linking Review

This subsection presents the review results in our considered dimensions, i.e., entity linking definitions, tasks, types, domains, approaches, datasets, applications, and challenges. Table 4 shows an overview of the occurrence frequency statistics for types, domains, applications, and challenges.

3.1.1 Entity Linking Definitions. We found 37 definitions of entity linking in the reviewed publications, which we describe hereafter and synthesize into the classification shown in Figure 4.

All but one of the definitions we found describe entity linking as an *end-to-end* process spanning several subtasks (see Table 5 and Figure 4). Delpeuch defines entity linking as: "the task of detecting mentions of entities from a knowledge base in free text" [12], extending from entity mention

[7]Table 10 in the Appendix shows an overview of the full results as a numeric frequency count across the entity linking dimensions of the publications considered in this review. For the full literal table, see https://zenodo.org/records/17839768.

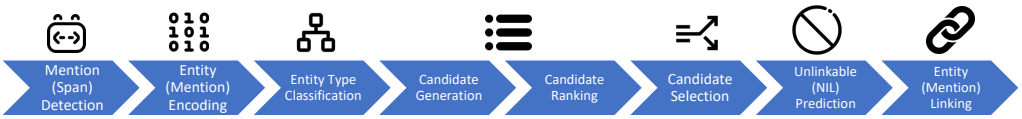


Fig. 4. Sythensized entity linking pipeline / workflow: The steps shown are applied to review entity linking definitions and tasks.

Table 5. Definition of Entity Linking Tasks with Sources

Task	Definition	Sources
End-to-end entity linking	entity recognition and disambiguation	[66]
Mention (span) detection	identifying mentions of named knowledge base entities in texts	[12, 29, 56]
Entity (mention) encoding	entity feature extraction and word embedding	[30, 58, 63, 82]
Entity type classification	classifying (named) entities by type (e.g., person, organization, location)	[30, 30]
Candidate generation	constructing an entity candidate list	[56, 63, 85]
Candidate ranking	ranking entities according to their confidence	[36, 70, 70]
Candidate selection	selecting the predicted entity from the ranking	[56, 82]
Unlinkable (NIL) prediction	prediction whether a given entity is absent in the knowledge base, i.e., "Not In Lexicon" (NIL)	[14]
Entity (mention) linking	linking mentions of knowledge base entity surface forms	[38, 51, 79]

detection to linking. Sakor et al. define entity linking as: "aligning unstructured text to its structured mentions in various knowledge repositories" [66]. The authors mention Wikidata along with Wikipedia, DBpedia, and Freebase as knowledge base examples. Both definitions are comparably high-level. Similarly, the entity linking definition of Mulang et al. includes the subtasks of entity recognition and entity linking. As in Delpeuch [12], the definition entails an end-to-end process spanning from recognition mention detection to linking. Sorokin et al. present an application-driven definition that positions entity linking as the: "first stage for every question answering approach" [75], focusing on identifying and linking entity mentions in question texts to a knowledge base.

Lin et al. focus their definition on short text, on which a combination of entity linking and relation linking can be applied to identify nouns and predicates [40]. Relation linking denotes identifying and mapping relations in sentences or queries, e.g., in the phrase "Who is the father of Barack Obama?" the relationship "father of" (Wikidata property P22) should be extracted. Banerjee et al. include the tasks of mention detection, entity disambiguation, candidate node selection, and entity linking to a knowledge base or knowledge graph [3]. Other authors describe the entity linking process as involving tasks like surface form extraction [51, 70, 71], i.e., mention detection [12], stance detection [26], named entity disambiguation [9, 46, 51, 56, 58, 59], contextualization [10, 76], classification [26, 61], embedding [36], candidate generation [56] and selection [20], Not In Lexicon prediction [14], and knowledge base or knowledge graph grounding [27, 85] or linking [6, 7, 9, 16, 34, 42, 51, 56, 82, 87, 89].

We observed that 16 of the definitions include *mention (span) detection*. An entity mention refers to the appearance of any of its surface forms within a text, while its span indicates the count and positions of the words involved. Delpeuch describes entity linking as: "the task of detecting mentions of entities" [12], identifying mention detection as the initial step in the entity linking process. Sakor et al. point out that a entity recognition approach: "aims to identify entity labels (or surface forms) in an input sentence" [66], where label identification is akin to mention detection, potentially encompassing the entire process from detection to linking. Mulang et al. define entity linking as: "concerned with the identification of entity surface forms in the text" [52]. Banerjee et al. define mention detection as the: "[identification of] a span of interest" [3]. Mulang et al. describe mention detection as surface form extraction, where a surface form is a: "contiguous span of text

that refers to a named entity" [51]. Other authors describe mention detection as identifying: "the relevant mention boundaries given a definite set of entity types" [9, 10], extracting mentions in an input sentence [66] or documents [70, 71], or locating named entities [26].

Only one of the reviewed publications explicitly incorporates *entity (mention) encoding* in its definition of entity linking. We still include this step due to its frequent application in neural entity linking [69]. Labusch et al. describe utilizing an: "[Approximative Nearest Neighbor] index that stores BERT-embeddings" and comparing Wikipedia text candidates through a: "purpose-trained BERT model" [36].

Two definitions of entity linking we encountered encompass *entity type classification*. Provatorova et al. state that they: "consider both named entity recognition and classification and entity mention detection tasks as instances of the sequence classification task" [61]. The definition of Hamdi et al. includes: "categori[z]ing [entities] into a set of pre-defined classes (i.e., person, location, organization, etc.)" [26].

Two definitions of entity linking we found include *candidate generation*. It is presented as: "candidate entity generation" [14] or specified as the construction of a "list of possible candidate [sic] for the identified entities" [56]. Only one entity linking definition in the publications we reviewed includes *candidates ranking*. Labusch et al. mention the: "final ranking of candidates based on information gathered from previous steps" [36].

We found that 17 definitions of entity linking in our review encompass *candidate selection or disambiguation*. Sakor et al. describe named entity disambiguation as the second entity linking sub-task, focusing on: "linking surface forms to semi-structured knowledge repositories" [66]. Mulang et al. extend this definition to include "structures" in knowledge bases [52], likely referring to knowledge base data structures. Other authors conceptualize candidate selection as choosing from a list of candidates [56], selecting appropriate candidates [20], contextualizing mentions [9, 10], or "matching a raw mention to the concept it references" [43]. Several authors mention named entity disambiguation without delving into the specifics of the disambiguation process [5, 7, 14, 26, 29, 36, 46, 51, 58, 59]. Within the publications we review, only one entity linking definition includes *unlinkable (NIL) prediction*. Pinto defines that if an entity mention score is below a threshold, "then the target entity of [this mention] is **Not In Lexicon** (**NIL**)" [14].

A considerable number, 22, of the entity linking definitions we encountered include *entity (mention) linking*. Several authors define entity linking as "linking [entities] to knowledge bases" [52, 66, 75], a definition that can refer both to the overall entity linking process and the specific step of entity mention linking. Lin et al. merge entity linking with to identify nouns and predicates [40]. Other authors detail entity mention linking as the process of connecting "[a] span of interest to the appropriate entity in the knowledge base or appropriate node in the knowledge graph" [3, 42], "the identified named entity to ground truth entities in a given knowledge base" [51], "entities to a distinctive identifier within a knowledge graph" [56], "[a] mention to the relevant entity in a [Knowledge Base]" [9, 70, 82],'mentions of uniquely separable things (which we can identify by a name, i.e. named entities) to unique identifiers" [27, 34], "mentions (surface names) in text to their corresponding entities in a Knowledge Graph" [14, 71], "[a] set of entities to pages from the underlying document collection that provide the context of their co-occurrence" [76]

Additionally, some authors focus on identifying "[an] entity's corresponding entry in a Knowledge Base" [6], selecting "suitable entity candidates [...] from the underlying Knowledge Base" [20], mapping [30, 89], grounding [85], or "assigning to parts of text (tokens) a unique identifier [(URI)] that points univocally to the referred entity in a given Knowledge Base" [7].

Our findings show that terms like recognition, identification, and detection are often used interchangeably in the literature. Similarly, entity labels, mentions, and surface forms are frequently used synonymously to denote the "contiguous span of text that refers to a named entity [51]."

3.1.2 Entity Linking Tasks. The reviewed literature describes the tasks of retrieving, detecting, identifying, classifying, ranking, disambiguating, evaluating, linking, and grounding entities.

We found 36 entity linking task descriptions in the reviewed publications. By consolidating these descriptions, we find that entity linking involves a combination of entity recognition (mention span detection) and disambiguation (encompassing candidate generation, ranking, and selection). It is important to note that not all steps in the framework are performed by every system. Sometimes, some of them are combined, named differently or even omitted. Hereafter, we describe Wikidata entity linking tasks using the same classification schema as for the entity linking definitions. Table 5 displays a description of the individual tasks that are summarized among the different definitions provided by the given sources.

Our results show that 25 publications discuss *end-to-end* entity linking. Sakor et al. outline two key sub-tasks: Named entity recognition and entity disambiguation. They define entity recognition as identifying: "entity labels (or surface forms) in an input sentence", while disambiguation involves "linking [them] to semi-structured knowledge repositories" [66]. Collectively the two tasks span the entire end-to-end entity linking process from identification to linking.

Additionally, the authors explore various tasks related to linking and disambiguating entities to knowledge bases and knowledge graphs. These tasks include automatically finding and linking ungrounded [6] surface forms [45, 52] or mention spans [3] of things [12, 16, 26, 34, 59, 71, 79, 89] to unique identifiers [27, 34] of their corresponding referents [38] or relevant ground truth [51] entities in a language-agnostic [79] knowledge base [3, 9, 32, 45]. We also identify task, such as recognizing (named) entities in free [12] text [7, 10, 16, 26, 29, 45, 52, 71, 79, 89] (e.g., noun phrases [38, 40], questions [75] or identifying pieces of text that refer to entities [59], and disambiguating with corresponding entities in knowledge graph [29]. Furthermore, end-to-end entity linking includes detecting, classifying, and linking (named) entities to enable semantic search [61] as well as person name disambiguation and linking techniques to identify the corresponding real-world entity for a person name [19].

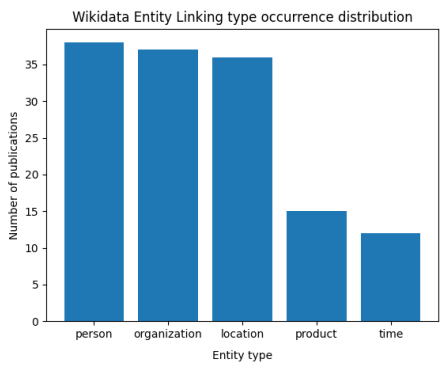
Seven publications discuss the *mention (span) detection* task. The task is described in various terms such as "detecting mentions of Knowledge Base entities in free text" [12], "identifying entities" [56], "recognizing (named) entities in text" [29], "detecting (named) entities" [61], "identifying pieces of text that refer to entities" [59], or simply "mention detection" [39, 70].

Four publications refer to the *entity (mention) encoding* task, either as "feature extraction" [63, 82] or entity and word embedding [30, 58]. Only two publications include descriptions of the *Entity Type Classification* task. Provatorova et al. mention "classifying (named) entities" [30], while Tempelmeier and Demidova phrase the task as "link classification" [30].

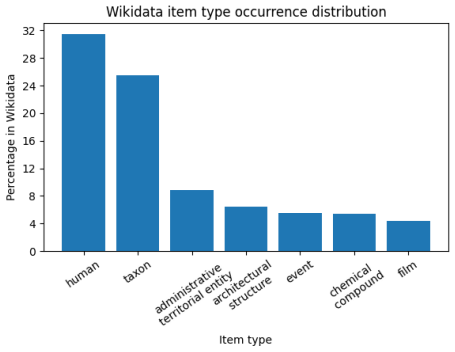
Six publications in our collection discuss the *candidate generation* task. Perkins highlights that entity linking involves constructing a candidate list [56]. Other authors alternatively term the task "candidate entity generation" [63, 70, 82], or simply 'searching' [85]. Only two publications [36, 70] explicitly mention the *candidate ranking* task. Shanaz and Ragel refer to candidate ranking as "candidate searching" [70].

Ten publications describe the *candidate selection and disambiguation* task. Several authors stress that entity linking involves disambiguating [39, 56] or selecting [82] (named) entities [5], candidates [85], or "surface forms of entity mentions in text" [45] and corresponding entities in a knowledge base or knowledge graph [29]. Geiß and Gertz focus on "person name disambiguation [19]". Other authors phrase the task as candidate selection [70] or candidate filtering and match correction [58]. Only one publication [14] discusses the *Unlinkable (NIL) Prediction* task.

The task of *entity (mention) linking* is discussed in 19 of the publications we reviewed. The authors describe entity linking as the process of linking (named) [38, 51, 61] entity surface forms [45, 52, 71]



5a



5b

Fig. 5. Entity type distributions. Figure 5(a) shows entity types in the reviewed publications, while Figure 5(b) shows the distribution of item types in Wikidata.

Table 6. Consolidation of Entity Type Categories Found in the Reviewed Publications

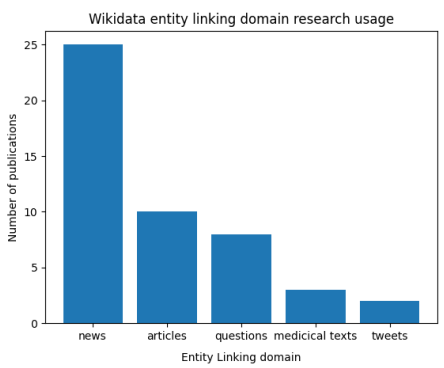
Type	Consolidations
person	actor, fictional character, cardinal, patient, publisher, developer
organization	band (music), company, institution, business, unincorporated community, author-affiliation, league, agency, political group, organisation
location	mountain, city, train station, geopolitical entity, country, municipality, airport, hospital, human settlement, village, river, communes of France, ocean, lake, administrative area, state, facility, transportation infrastructure, building, moon, volcano, skyscraper, mountain, cathedral, place, castle
product	album (music), film, single (music), literary work, television series, song, software, video game console, painting, work, human product
time	date, event, season (sports)

of real-world [19] things [34], persons [19], or mention [3, 7, 9, 26, 32, 34, 38, 45, 75, 79, 89] spans [3], e.g., in documents [7] or pieces of [59] text [59, 71], e.g., noun phrases [38, 40], to relevant [9, 32] ground truth [51, 85] data structures [52] or identifiers [56] or corresponding [16, 19, 26, 71] referents [16, 59] in a language-agnostic [79] knowledge base or knowledge graph.

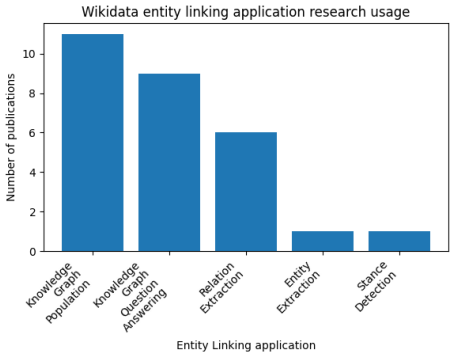
In summary, most publications describe the holistic end-to-end entity linking process. In addition, the authors discuss the subtasks for linking entities, such as entity recognition, named entity detection, named entity recognition and detection, entity detection, entity recognition, mention detection, and stance detection. In summary, authors conceptualize entity linking as the process of aligning unstructured text with structured entity mentions in a knowledge base.

3.1.3 Entity Linking Types. Figure 5(a) shows the distribution of the five most frequent entity types in the reviewed literature and the number of publications for each type.

To compute statistics, the diverse set of entity types was consolidated (see Table 6) into commonly adopted primary categories (according to Ref. [22]): person, organization, location, product, and time. This involved mapping related terms (e.g., actor and fictional character to person, company and agency to organization, city and country to location, film and software to product, and date and event to time) without altering the entity type order. A full list of all entity types and their frequencies is available in the Appendix.



6a



6b

Fig. 6. Entity linking domains (6(a)) and entity linking applications (6(b)) most frequently adressed in the reviewed publications.

The frequencies of entity types occurring in the reviewed publications (Figure 5(a)) were compared with the frequencies of item types in Wikidata (Figure 5(b)). Wikidata statistics [8] indicate that the person type (here "human") constitutes only about 9% of items, while organization and location types are even less prevalent. This discrepancy highlights a structural misalignment between Wikidata's data and academic research needs, limiting its utility for certain studies or requiring translation between multiple item types. By consolidating entity types and analyzing their distribution, this study underscores the importance of aligning entity linking resources with research objectives for enhanced applicability and usability.

3.1.4 Entity Linking Domains. We identified 44 entity linking domains (i.e., fields or use cases where entity linking is applied) from the reviewed publications. Figure 6(a) illustrates the five domains that occur most frequently in the literature. We describe the most cited publications within these five examples of Wikidata entity linking domains, ordered by the frequency of domain usage in the reviewed publications. As for the entity types, a full list of all entity linking domains and their occurrence frequencies in the reviewed publications can be found in the Appendix.

The entity linking domain *news* is featured or discussed in 20 of the reviewed publications. Within these, 13 publications focus on "general news", six on "historical news", and one on "economic newspapers". We deduce the general news domain for the respective publications based on the domains of the datasets they utilize, such as the AIDA CoNLL-YAGO dataset [28]. In the following, we describe some dataset examples. The NYT2018 includes 30 news documents, each manually annotated with links to Wikidata and DBpedia ressources [50]. The AIDA CoNLL-YAGO dataset [28] originates from the "CoNLL" 2003 [9] shared task [67] utilizing a corpus of Reuters news stories. The Mewsli-9 dataset (Multilingual Entities in News, linked) allows for multilingual entity linking covering 100 languages [6]. We infer the historical news domain for two publications [16, 26] as the domains of their employed datasets, e.g., the CLEF HIPE 2020 dataset [59]. Boros et al. demonstrate multilingual entity recognition and entity linking on historical multilingual documents [5] from the CLEF HIPE 2020 task [16]. Hamdi et al. present a multilingual dataset, NewsEye, for entity recognition, entity linking, and relation extraction in historical newspapers [26].

[8] https://www.wikidata.org/wiki/Wikidata:Statistics
[9] http://lcg-www.uia.ac.be/conll2003/ner

The entity linking domain *articles* is featured or discussed in eight of the reviewed publications. Among these, three focus on "Wikipedia articles", one on "Wikipedia abstracts", three on "encyclopedias", and one on "research articles". We infer the domain of considered publications that present entity linking on Wikipedia articles by their employed datasets. For example, Mulang et al. use "T-REx",[10] a dataset for aligning natural language with knowledge base triples (between DBpedia abstracts and Wikidata triples), containing 11 million triple alignments from 3.09 million DBpedia abstracts (6.2 million sentences) [52]. Labusch et al. [36] test their entity linking approach, including **Optical Character Recognition** (**OCR**) with BERT, on a dataset derived from the German Wikipedia, focusing on identifying persons, locations, and organizations.

The "Kensho Derived Wikimedia Dataset" [56] was released in 2020 by Kensho R&D group, comprising an English Wikipedia corpus and Wikidata knowledge graph for entity linking and other natural language processing tasks. Mesquita et al. [11] present KnowledgeNet, a benchmark dataset for knowledge base population containing DBpedia abstracts (i.e., first paragraphs of a Wikipedia page, what can be regarded as a "Wikipedia abstract"). Zhou et al. release Richpedia-MEL, a dataset built from a multimodal Wiki knowledge graph, featuring textual and visual descriptions for over 17K multimodal samples and 20,752 mention-entity pairs [89]. Weichselbraun et al. use two datasets, covering encyclopedic short texts with descriptions of subjects (e.g., short DBpedia abstracts), focusing on PER, ORG, and LOC entity types [85]. Lin et al. assess their TENET approach for joint entity and relation linking with coherence relaxation on the "T-REx42" dataset (long-text, 179.17 words/document), a benchmark for knowledge base population, relation extraction, and question answering [17]. We infer the domain of the considered publication that presents entity linking on Wikipedia articles as the domain of the employed dataset. In this case, Delpeuch proposes "OpenTapioca" [12] a simple named entity linking system, lightweight to train on Wikidata, easy to run and keep synchronous with Wikidata in real time. The approach is evaluated using the ISTEX dataset, comprising 1K author affiliation strings extracted from research articles, provided by the ISTEX text and data mining service.[11]

The entity linking domain *questions* is featured or discussed in seven of the reviewed publications. Of these, six publications address "web questions", while one focuses on "general complex questions". We deduce the web questions [12] domain from the domains of the datasets the authors employ for their experiments, such as the SimpleQuestions dataset [4]. For example, the approach by Huang et al. that performs entity linking for short text using a structured knowledge graph and multi-grained text matching employs the WebQSP dataset [13] that contains 5,810 questions, partially with annotated SPARQL queries [88]. Dubey et al. present LC-QuAD 2.0, a large dataset for general complex question answering over Wikidata and DBpedia, consisting of 30,000 questions, their paraphrases, and corresponding SPARQL queries [15]. Lin et al. validate the performance of their system "KBPearl" for knowledge base population from unstructured text [40] on the question answering baselines LC-QuAD 2.0[15] and QALD-7-Wiki.[14] Banerjee et al. evaluate their end-to-end entity linking system over knowledge graphs for question answering [3] on the datasets WebQSP,[13] SimpleQuestions [4], and LC-QuAD 2.0 [15]. Liu et al. evaluate entity linking over question answering pairs with structured triples (head entity, relation, tail entity) on their own dataset created from the "Baidu Knows" website [15] (HTML files) [42].

[10] https://hadyelsahar.github.io/t-rex
[11] https://www.istex.fr
[12] Questions posed on the web, i.e., using the web as a knowledge base for answering complex questions.
[13] https://paperswithcode.com/dataset/webquestionssp
[14] https://qald.aksw.org/
[15] https://zhidao.baidu.com/

The entity linking domain *medical texts* is featured or discussed in three of the reviewed publications. Michel et al. introduce Covid-on-the-Web, a knowledge graph and services designed to advance COVID-19 research by aiding biomedical researchers in searching and understanding relevant literature. To achieve this, they combine various approaches to analyze and enrich the "COVID-19 Open Research Dataset" (CORD-19), which comprises over 50,000 articles. This dataset includes two knowledge graphs: (1) named entities mentioned in the CORD-19 corpus, linked to DBpedia, Wikidata, and other BioPortal vocabularies, and (2) arguments extracted using ACTA, a tool for automation, extraction, and visualization of argumentative graphs. This tool aids clinicians in analyzing clinical trials for decision-making purposes [46].

Lopez et al. utilize the "**PubMed Central** (**PMC**) Open Access Subset" dataset [16] from the medical domain. They incorporate this dataset into the newly developed "Softcite" dataset, which focuses on software citations. It includes 8,336 annotations of software names and attributes (version number or date, publisher, URL) from 4,971 randomly selected full-text research articles. The annotations were produced through a multi-round human labeling task involving 38 annotators and two meta-reviewers, achieving a 75.5% inter-annotator agreement rate [43].

Schindler et al. present the "SoMeSci" dataset,[17] aimed at software mentions in articles from medicine, biology, life sciences, and related disciplines. They highlight the lack of formal software citations in research articles, with a preference for informal mentions, underscoring the need for automatic information extraction and disambiguation methods. The SoMeSci benchmark dataset comprises curated annotations of 3,756 software mentions (including name, version, developer, URL) in 1,367 PubMed Central medicine-related articles [68].

The entity linking domain *tweets* is featured or discussed in one of the reviewed publications. Haradizadeh and Singh present "Tweeki", an unsupervised, modular method for linking entities in Twitter posts to the Wikidata knowledge graph. The authors introduce two tweet datasets: "TweekiData", which is automatically annotated, and "TweekiGold", a benchmark dataset [27]. Both datasets are designed to be used to enhance downstream tasks in social media analysis, such as geolocation prediction.

3.1.5 Entity Linking Approaches. We identified 34 entity linking approaches in the publications we reviewed. By approaches, we denote the combination of the individual publication's presented systems together with their employed methods. We describe the five most frequently cited and discussed approaches for Wikidata entity linking as representative examples.

Delpeuch presented "OpenTapioca",[18] an entity recognition system trained exclusively on Wikidata. This approach highlights Wikidata's advantages and limitations as a data source, offering a reproducible baseline for comparative evaluations with other sources and methods. The system combines surface-form matching and graph-based semantic similarity to identify and disambiguate entities in text. The model leverages features such as log-linear estimations of entity popularity and proximity-based contextual coherence, using a Markov chain to propagate local evidence across a semantic graph of candidate entities. Additionally, OpenTapioca uses curated surface forms and real-time synchronization with Wikidata to maintain high-quality and up-to-date entity mappings without relying on external data sources [12].

Sakor et al. introduced "Falcon,",[19] a tool for joint entity and relation linking over Wikidata. The public system employs N-Gram tiling and splitting to process short English texts into a ranked list of entities and relations annotated with their Internationalized Resource Identifier (IRI) in the

[16] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist
[17] https://data.gesis.org/somesci/
[18] https://github.com/opentapioca/opentapioca
[19] https://labs.tib.eu/falcon/falcon2

Wikidata knowledge graph. Falcon leverages a rule-based linguistic approach grounded in principles of English morphology, such as tokenization and compounding, to identify surface forms for entities and relations. These surface forms are matched against a background knowledge base enriched with aliases and synonyms extracted from Wikidata, allowing for robust linking performance even in noisy or ambiguous contexts. The system also integrates candidate ranking and verification processes using an RDF triple store to ensure accurate matches. The authors offer convenient access to Falcon through an API, making it a practical tool for researchers and developers working on natural language processing and semantic web applications [66].

Mulang et al. propose an attentive neural network model that incorporates the knowledge graph context as background knowledge to tackle the entity linking challenge posed by the often non-standard, noisy, and lengthy entity titles in collaborative knowledge graphs like Wikidata. This issue can diminish performance in terms of precision and recall. Their approach demonstrates superior performance over baseline models and other end-to-end entity linking systems specifically tailored for Wikidata [52].

Sorokin et al. presented a neural architecture specifically designed for detecting and disambiguating entity mentions within question-answering contexts. This method leverages the surrounding context of entities at varying granularity levels, optimizing them jointly. The authors evaluated their system on several question-answering benchmark datasets and observed robust performance across diverse entity categories [75].

The "KBPearl" system represents an end-to-end approach for knowledge base population from unstructured text. It employs joint entity and relation linking to augment and populate an incomplete knowledge base using contextual knowledge and side information extracted from a vast corpus. KBPearl organizes the noisy data obtained from open information extraction into canonicalized facts. Through comprehensive experiments on real-world datasets, the authors illustrate the system's effectiveness and efficiency, demonstrating that KBPearl surpasses contemporary state-of-the-art knowledge base population techniques [40].

3.1.6 Entity Linking Datasets. We identified 17 entity linking datasets from the reviewed publications. Hereafter, we describe the (top) five most commonly used (employed in the experiments of the respective papers) Wikidata entity linking datasets. Note that it is outside the scope of our study to catalog and compare the individual statistics of the datasets. We refer to Moeller et al.[50] for this type of analysis.

Dubey et al. present "LC-QuAD", a large, complex question-answering dataset [20] containing 30,000 questions, paraphrases, and corresponding SPARQL queries. The dataset is compatible with the Wikidata and DBpedia knowledge graph. It aims to advance research in translating natural language questions into formal queries, enabling machines to navigate knowledge graphs and provide answers. LC-QuAD 2.0 surpasses its predecessors like WebQuestions, QALD, and its earlier version by offering a broader variety of questions and a larger scale. The publication details the dataset's creation process, showcases question examples, and presents a statistical analysis of the data [15].

Ehrmann et al. introduce the "HIPE" dataset, comprising historic newspapers.[21] The dataset was developed to identify **HI**storical **P**eople, places, and other **E**ntities, serving as a shared task [22] in entity recognition and entity linking within multilingual historical documents. The authors evaluate Named Entity processing in documents written in French, German, and English. HIPE aims to enhance the robustness of existing approaches to non-standard inputs, facilitate the comparison

[20] https://sda.tech/projects/lc-quad-2/
[21] https://impresso.github.io/CLEF-HIPE-2020/datasets.html
[22] https://hipe-eval.github.io/HIPE-2022

of NE processing performance in historical texts, and promote efficient semantic indexing of historical documents. The publication details the tasks, corpora, and results from 13 participating teams. It also provides additional information on data generation, statistics, and the systems used [16].

The "T-REx" dataset,[23] a large-scale collection of alignments between Wikipedia abstracts and Wikidata triples, is crucial for training machine learning models used in various natural language processing tasks. Other datasets often have limitations, such as small size, limited predicate coverage, and unreported quality. T-REx overcomes these issues by offering 11 million triples aligned with 3.1 million Wikipedia abstracts. This makes T-REx two orders of magnitude larger than the previously largest available dataset and provides 2.5 times more predicate coverage. The dataset's quality is established through extensive crowdsourcing evaluation. The publicly accessible dataset is particularly useful for tasks like relation extraction, knowledge base population, question answering, and natural language generation from knowledge graph triples [17].

Hoffart et al. present the "AIDA CoNLL-YAGO" dataset,[24] designed for "Robust Disambiguation of Named Entities in Text". This dataset accompanies a robust method for disambiguating named entities in natural language texts, leveraging context from knowledge bases, and employing a novel coherence graph. The approach integrates prior methods into a comprehensive framework, combining three key measures: prior probability, context similarity, and coherence among candidate entities. It constructs a weighted graph linking mentions and candidate entities, then computes a dense subgraph to approximate the optimal joint mention-entity mappings. The dataset provides assignments of named entities to their respective knowledge base URLs, along with the mention-entity candidate mapping utilized in their experiments [28].

Finally, the "SimpleQuestions" dataset,[25] was introduced by Bordes et al. It contains 100,000 open domain questions and is tailored to evaluating multitask and transfer learning in large-scale simple question-answering scenarios. The primary challenge of this dataset lies in its limited training resources, which only cover a fraction of the possible spectrum of questions. In their experiments using Memory Networks, Bordes et al. demonstrate that these networks can be trained effectively to achieve remarkable performance. The authors propose that this strategy is a stepping stone towards scaling up to more intricate forms of reasoning [4].

The review of entity linking datasets reveals significant progress in aligning natural language with structured knowledge bases, leveraging diverse data sources and methodologies. These datasets are designed to address challenges such as question answering, named entity recognition, and relation extraction across domains such as historical texts, large-scale knowledge graphs, and limited-resource environments. They demonstrate advances in scale, multilingual capabilities, and the ability to handle complex queries and non-standard inputs. The insights highlight the importance of dataset quality, with innovations, such as crowdsourcing evaluations and novel modeling frameworks ensuring robustness and applicability. The reviewed datasets collectively push the boundaries of semantic understanding and reasoning in AI, facilitating improved performance in linking textual data to structured representations, and driving advancements in knowledge-driven natural language processing.

3.1.7 Entity Linking Applications. We identified 23 entity linking applications from the reviewed publications. Figure 6(b) shows the five applications most frequently discussed in publications we review. These five applications are presented hereafter. A complete list of entity linking applications is found in the Appendix.

[23] https://w3id.org/t-rex
[24] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/aida/downloads
[25] https://paperswithcode.com/dataset/simplequestions

Knowledge base population or knowledge graph population are discussed in eight of the reviewed publications, with six focusing on knowledge graph population and two on knowledge base population. Lin et al. utilize knowledge base population to enrich a knowledge base by jointly linking entities and relations, integrating facts and additional information extracted from source documents [40]. ElSahar et al. employ the TAC-KBP knowledge base population dataset for their research [17]. Mesquita et al. introduce "KnowledgeNet," a benchmark dataset designed for the knowledge base population task of automatically enriching Wikidata with facts sourced from natural language texts on the web [11]. González et al. leverage entity linking as a means of populating skill ontologies [21]. Luggen et al. develop statistical approaches for estimating class cardinalities in collaborative knowledge graph platforms, contributing to efforts aimed at achieving completeness in knowledge base population [44]. Metilli et al. [45] focus on populating narratives using Wikidata events. In their knowledge graph population experiment, they create or enhance a "**Wikidata Event Graph**" (**WEG**), a graph representing implicit events identified in Wikidata, such as the inference of a birth event from a listed date of birth. Portisch et al. devise methods for analyzing and mining the "COVID-19 Open Research Dataset" (CORD-19) to construct an argumentative knowledge graph for medical research to combat the COVID-19 pandemic [60].

Question answering over knowledge graphs is featured or discussed in seven of the reviewed publications. Sakor et al. introduce Falcon 2.0,[26] a joint entity and tool designed for Wikidata. They evaluate their tool on short text questions, employing a catalog of rules, such as using a question's headword (who, where, when, etc.) to resolve ambiguities by filtering for persons, locations, dates, and so on. [66]. Sorokin and Gurevych assess entity linking in question answering tasks and present a neural architecture that is jointly optimized for entity mention detection and entity disambiguation. This system models context at varying levels of granularity and offers a benchmark for entity linking in question answering data [75]. Banerjee et al. evaluate a pointer network-based end-to-end entity linking system over knowledge graphs for question answering over knowledge graphs applications. The system operates without querying the Wikidata knowledge graph during runtime, using pre-indexed entity labels, descriptions, and encodings to answer questions [3].

Dubey et al. develop LC-QuAD 2.0,[27] a comprehensive dataset for complex question answering that includes 30,000 natural language questions, their paraphrases, and corresponding SPARQL queries. This dataset is tailored to advance research in natural language question answering over knowledge graphs, particularly Wikidata and DBpedia. The dataset's creation and statistical analysis are also detailed [15]. Pinto and Alejandro observe that recent question answering over knowledge graphs approaches based on neural semantic parsing often adopt a neural machine translation style, translating natural language questions into structured query languages. To overcome the out-of-vocabulary issue (where terms in a question might not have been seen during training), they propose a question answering over knowledge graphs method that assigns the processing of entities to entity linking systems, resulting in a query template with entity placeholders. This combination of entity linking and neural semantic parsing demonstrates promising performance improvements in question answering over knowledge graphs tasks over Wikidata [14].

Sorokin and Gurevych also introduce a knowledge graph question answering system for the QALD-7 shared task [28] that features an end-to-end neural architecture for the stepwise construction of a structural semantic knowledge base query from a natural language question. They employ a

[26] https://labs.tib.eu/falcon/falcon2
[27] http://lc-quad.sda.tech
[28] https://github.com/UKPLab/eswc2017-question-answering

convolutional neural network to learn vector encodings for questions and semantic graphs, aiding in selecting the best graph for the input question [74]. Liu et al. evaluate entity linking over question answering pairs where the question and answer entities are semantically related but not identical. By analyzing linked factual triples, they mine global knowledge, such as the likelihood of relations and the linking similarity between question and answer entities [42].

Relation extraction is featured or discussed in five of the reviewed publications. The Falcon 2.0 approach of Sakor et al. for joint entity and relation linking over Wikidata transforms short English text into ranked lists of candidate entities and relations. In this context, the extracted relations can be Wikidata properties like death place' or spouse'. The authors report that question answering and linking are only half as effective (as measured by F-score) as entity extraction and linking [66]. Lin et al. focus on populating the Wikidata knowledge graph by joint entity and relation linking (also known as "slot filling"), using facts and side information extracted from source documents. Their end-to-end system, KBPearl, does not require preliminary specifications of predicates of interest to retrieve canonicalized triples [40].

ElSahar et al. tackle the issue of restricted relation predicate coverage by introducing T-REx,[29] a dataset featuring large-scale alignments between 3 million Wikipedia abstracts and 11 million Wikidata triples. This dataset surpasses previous state-of-the-art baselines by covering 2.5 times more predicates (for relation extraction) with high quality, thanks to extensive crowdsourcing evaluation [17]. Yang et al. developed a "**Relation Linking System for Wikidata**" (**RLSW**) that clusters relation mentions in text using a novel phrase similarity algorithm. This system also utilizes word location information and employs "a bag of distribution pattern modeling method" [87]. Schindler et al. present SoMeSci (Software Mentions in Science), a gold standard knowledge graph of software mentions in scientific articles, designed to jointly aid entity recognition, relation extraction, and entity detection tasks. This dataset comprises 3,756 software mentions in 1,367 PubMed Central articles. For relation extraction tasks, the publishers also provide relation labels for additional property information, such as version, developer, and programming environment [68].

Entity extraction is discussed in one of the reviewed publications. Pontes et al. focus their entity linking research on historical documents, such as newspapers and letters. As discussed in Section 3.1.8, this domain presents significant challenges due to language variations and OCR errors. The authors carry out EE in multiple European languages (English, Finnish, French, German, and Swedish) to demonstrate that their system enhances the overall performance (F-score) across all languages and included datasets [58].

Stance detection is discussed in one of the reviewed publications. Hamdi et al. analyze fine-grained searches on **Open Government Data** (**OGD**) to create links from OGD portals and catalogs to the Wikidata knowledge graph. Their multilingual dataset, covering German, French, Finnish, and Swedish, includes annotations for relation extraction of given entities as a sequence pair classification task. The sequence involves (1) the body text, (2) the identified entity, and (3) the class label, which can be positive, negative, or neutral. The authors report F1 scores reaching up to 0.579 for the German language. In conclusion, they outline plans to further refine guidelines for relation extraction to enhance the quality of the suggestions and explanations provided [26].

3.1.8 Entity Linking Challenges. We identified 216 challenges related to entity linking from the reviewed publications, where they are described as "challenges," "issues," or "problems," typically in the "Introduction" or "Future Work" sections. To ensure relevance, we excluded four challenges that referred only to adjacent topics, such as the knowledge base population or triple extraction. This section synthesizes the five most frequently discussed Wikidata entity linking challenges. We

[29] https://w3id.org/t-rex

ranked the challenges according to the number of papers that mentioned them. We merged closely related topics, such as those concerning knowledge bases and sparsity, to create a cohesive analysis. Descriptions and examples of these challenges are detailed in the Appendix.

In summary, our analysis reveals a key demand for *specific approaches* tailored to complex problems, highlighted in 18 of the reviewed publications. For example, Sakor et al. identified a gap in hybrid approaches that integrate rule-based and machine learning techniques [66]. Similarly, Lin et al. emphasized the importance of addressing joint entity and relation linking [40]. These methods promise to bridge the gaps in traditional approaches by leveraging multiple methodologies for better precision and adaptability. Perkins proposed the use of graph algorithms to enhance candidate generation by incorporating a deeper semantic understanding of text [56]. Provatorova et al. similarly endorsed graph-based disambiguation methods and suggested refining Transformer models with advanced architectures, parameter tuning, and noise reduction techniques such as OCR correction [61]. These innovations highlight the importance of integrating advanced algorithms and fine-tuning to improve the robustness of entity linking systems. Botha et al. discussed the challenges of (re)ranking candidates for long-tail entities and non-English languages, advocating for more sophisticated approaches to address these shortcomings [6]. Furthermore, Canale et al. highlighted the lack of standardization among named entity recognition systems, which hampers tunability, customization, and cross-model comparisons [7]. This underscores the need for standardized frameworks that can accommodate diverse requirements and contexts.

A recurring theme in the reviewed literature is the need for transparency and explainability in neural methods. Ilievski et al. emphasized that while neural approaches have advanced performance, it remains unclear which aspects of knowledge they effectively capture [30]. Enhancing explainability in neural models could improve trust and usability in broader applications.

In summary, the challenges in entity linking are diverse and multifaceted, ranging from technical innovations in hybrid and graph-based methods to systemic improvements in standardization and explainability. Addressing these interconnected issues will require ongoing collaboration across the fields of Natural Language Processing, Semantic Web technologies, and Machine Learning.

3.2 RT 2. Entity Linking Framework

This subsection summarizes the main findings from the reviewed publications and systematically compiles definitions and (sub)tasks related to entity linking. Recall that our aim is to construct a comprehensive theoretical framework for the entity linking pipeline. This framework will integrate insights and methodologies from the reviewed literature, providing a structured approach to entity linking.

The *end-to-end process* of entity linking reaches from surface form extraction [51, 70, 71], i.e., mention detection [12], stance detection [26], named entity disambiguation [9, 46, 51, 56, 58, 59], contextualization [10, 76], classification [26, 61], embedding [36], and candidate generation [56] to candidate selection [20], Not In Lexicon prediction [14], and grounding [27, 85] or linking [6, 7, 9, 16, 34, 42, 51, 56, 82, 87, 89] entities with corresponding representations in a knowledge base or knowledge graph. The overarching task of entity linking is to map N-grams from a text collection to corresponding entity representations in a knowledge base or assign the label "NIL" when appropriate. Essentially, entity linking performs word-sense disambiguation through a knowledge base, resolving lexical ambiguities by identifying the specific meanings of entities within their context. Entity linking systems can recognize either coarse-grained entity types (such as Person, Organization, Location, etc.) or more fine-grained types, depending on their design and the requirements of the application [69, 72].

In *mention detection*, a system processes text input (sentences, questions, etc.) to identify and locate mentions or surface forms, i.e., labels of named entities [12, 26, 29, 39, 52, 56, 59, 61, 66].

The identification of a contiguous text span of interest that refers to a named entity is called span detection [3, 42, 51]. In both mention detection and span detection, the system must detect the relevant mention boundaries given a set of entities in a knowledge base [9, 10]. Therefore, it needs to compare and map different subsets of the text (the text spans) to the entity names represented in the knowledge base [30]. Mention detection and span detection can be considered as binary text sequence classification tasks [61], mapping N-grams of the text sequence to 1 if they include a given entity mention and 0 if not. Moreover, the tasks can be considered as information extraction tasks, i.e., extracting structured (named) entity sets from unstructured or semi-structured text [70, 71].

Entity (mention) encoding is the task of vectorizing entity representations. This includes feature extraction [63, 82], e.g., TF-IDF or BERT [13] or link distance, and entity (word or N-gram) embedding [30, 58], e.g., as a Wiki graph [69]. The resulting entity representations can be stored in an index to have faster access to possible entity linking candidates, e.g., for Approximative Nearest Neighbor retrieval [36]. Enhancing the generalization capabilities of entity linking systems and models involves focusing on mention-context encoding, which considers the semantic context surrounding an entity mention. This approach typically involves creating dense, contextualized vector representations that encapsulate the structure of a knowledge graph (entity relationships) or entity definitions, as well as textual information found in large, annotated corpora. These elements are encoded into low-dimensional vectors for efficient processing. Neural entity linking techniques include convolutional or recurrent encoders, LSTMs, tensor networks, as well as (self-)attention between candidate entity embeddings and embeddings of words surrounding a mention (concatenated and jointly encoded across all mentions in a coreference chain) [69].

Entity type classification, sometimes referred to as link classification [30], is a process in entity linking that categorizes entities identified in a text into pre-defined classes, such as person (PER), location (LOC), and organization (ORG), and so on, to categorize entity mentions [26, 61, 82]. The commonly used combination of named entity recognition and entity type classification is called named entity recognition and classification. Both entity mention detection (see previous paragraph) and named entity recognition and classification can be considered as sequence classification tasks [61].

(Entity) candidate generation [63, 70, 82] or searching [85] is the task of constructing a list or set of possible candidates (or "senses") for the identified entities or entity mentions [56]. The task of candidate generation is very challenging since a mention potentially can be linked to any entity in a knowledge graph, which commonly results in a very large search space that needs to be reduced [69]. This includes filtering out irrelevant candidate entities from the grounding knowledge base. Among the commonly used approaches are dictionary-based techniques, surface form expansion from the local document, string or N-gram matching, and methods based on search engines or other heuristics. Systems can draw information from entity pages, redirect pages, disambiguation pages, hyperlinks in Wikipedia articles, and "CrossWikis". Also, supervised (machine) learning methods, such as feature vector representations and classifiers on acronym-expansion pairs are employed [72]. Three prevalent models for candidate generation in neural entity linking rely on surface form matching (employing methods like Levenshtein distance, N-grams, and normalization), alias expansion, and prior probability computation. Often, these approaches are combined in candidate generation, including in recent zero-shot models that accomplish this task without relying on external knowledge [69].

In *candidate ranking*, the entity linking system sorts the list of candidates by relevance to identify or "search for" [70] the most pertinent entity in a knowledge base for a given mention in a text, i.e., the entity that is most likely being referred to. Thus, entity linking systems typically evaluate various factors informed by information gathered from earlier stages in the entity linking information extraction pipeline [36], including the context of the mention, the similarity between the mention

and the entity, and the entity"s popularity or authority. In some cases, candidate filtering, i.e., removing a set of irrelevant candidates from the list, precedes candidate ranking [58].

In *candidate selection or disambiguation*, the most relevant candidate is chosen [82] from the (ranked) list of candidates [70]. This process may involve match correction [58], which entails altering the selection if certain criteria are not fulfilled. Disambiguation [39, 56] via selection from a filtered and ranked list of entity grounding candidates is typically achieved through contextualization [9, 10], which involves considering the text surrounding the entity mention to be linked. One of the most prevalent forms of entity disambiguation is person name disambiguation [19], a crucial process due to the common occurrence of multiple individuals sharing the same name.

Unlinkable (NIL) prediction: In the entity linking process, the target NIL (referring to "Non-Informative Label" or "Not In Lexicon") is typically predicted as a placeholder to represent an entity that cannot be accurately identified or resolved to a specific concept in a grounding knowledge base. This may occur if an entity mention score is below a threshold [14], the entity mention is ambiguous, or when insufficient information is available to determine the correct entity to link. The entity linking system module typically examines the top-ranked entity or entity candidates to ascertain if a specific target is in the knowledge base, defaulting to "Not In Lexicon" otherwise. Simple heuristics may be employed, such as predicting NIL when the candidate entity set is empty [72]. The entity confidence threshold can be derived from training data or determined using a binary machine learning classifier to assess the likelihood of the top-ranked candidate being correct or not. Additionally, a NIL entity might be included in the candidate list prior to ranking or as a placeholder in the knowledge base.

Entity (mention) linking constitutes the final stage in the entity linking process. Note that, in the literature, entity linking is used to denote both the overall process [52, 66, 75] and the specific final step. The task involves mapping a selected candidate meaning for a detected mention (text span of interest) to an appropriate entity item in a knowledge base or node in a knowledge graph [3, 42]. This means assigning a **unique identifier** (**URI**) [7] to the text passage that points unambiguously to the referred entity in a given grounding knowledge base [85].

3.3 RT 3. Entity Linking Research Gaps

This section addresses Research Task 3 by outlining the research gaps within each of the six review dimensions.

3.3.1 Gaps Per Review Dimension.

Types. The prevalent entity types identified in works addressing entity linking with Wikidata as the target knowledge base include person (PER), location (LOC), organization (ORG), "product" (notable for its industrial relevance), and "time" (specified as "datetime"). However, the research predominantly emphasizes the primary types (PER, LOC, and ORG), with less attention to fine-grained or specialized types. This creates a gap in addressing diverse domains and applications. For instance, some authors [19, 30, 87, 89] focus on the entity type person (PER). Yang et al. use the category "HUMAN" as a synonym for *person* (PER) [87], while Ilievski et al. explore the use of side knowledge for detecting long-tail person entities [30]. These studies address challenges in identifying individuals across varying contexts but do not extend to related subtypes like fictional characters or historical figures. The organization (ORG) entity type is commonly used across studies [12, 26], often including subtypes such as "company", "institution", "agency", and "political group". Notably, no publications exclusively focus on organizations, despite their critical role in applications like stance detection and sentiment analysis [26]. The location (LOC) entity type encompasses a variety of subtypes, such as "city", "country", "mountain", "cathedral", or

"castle". Some studies employ methods like cosine similarity computations [71] or entity coherence assessments on selected datasets [28] to address location-related challenges. The product type (or "human product") is primarily used for media, such as "literary work", "single (music)", "album (music)", "television 'series', "film", or "video game". Despite its commercial relevance in applications like sentiment analysis for brands and products [26], this type receives less research focus compared with the primary categories. The entity type time includes related concepts, such as "date" and "event", and is often used for event extraction. For example, Spitz et al. employ implicit networks for entity exploration, summarization, and linking in event-related contexts [76].

In summary, the current literature predominantly focuses on the primary entity types: person (PER), organization (ORG), and location (LOC). Fine-grained types, such as product, time, and miscellaneous categories, are less common. To address this gap, future research should prioritize specialized entity types, such as music (e.g., songs or albums), or software citations, incorporating details like version, developer, and URL. Expanding the scope to these underrepresented types would significantly enhance the applicability of entity linking systems in diverse domains.

Domains. The domains most frequently addressed in the reviewed literature include *news* (current and historical), (web) *questions*, (Wikipedia and research) *articles*, *tweets*, and *medical texts*. While these domains align closely with the focus of existing entity linking approaches and datasets, there is a noticeable gap in exploring less conventional domains, such as mathematics or programming code, which could broaden the applicability of entity linking systems. The *news* domain is predominant, divided into general, historical, and economic news. Several datasets cater to this domain, including the NYT2018 dataset [50] and the AIDA CoNLL-YAGO dataset [28], which is based on the "CoNLL" 2003 [30] shared task [67] on Reuters news stories. Entity linking approaches for news include network-based topic extraction [77] and multilingual entity linking [5, 26]. These works address a range of news-related challenges but do not extend to emerging subdomains like niche news platforms or hyperlocal journalism. In the *questions* domain, several datasets have been developed, such as SimpleQuestions [4], WebQSP [88], and GraphQuestions [78]. A prominent approach involves entity linking over question-answering pairs with structured triples (head entity, relation, tail entity) [42]. However, the focus remains largely on structured data, leaving unstructured and conversational question-answering domains underexplored. The *articles* domain consists of Wikipedia articles, abstracts, and research publications. Notable approaches include evaluations of OCR with BERT [36] and joint entity and relation linking with coherence relaxation [17]. While these studies are valuable, they primarily target encyclopedic and general academic content, with limited emphasis on domain-specific literature, such as legal or financial texts. In the *tweets* domain, benchmarks like TweekiGold [27] have been developed to enhance downstream tasks in social media analysis, such as geolocation prediction. Despite these advancements, the focus on geolocation and sentiment analysis leaves room for expanding entity linking in areas like misinformation detection or crisis response on social media. The *medical* domain has seen limited but impactful contributions. Michel et al. combine approaches to analyze and enrich the "COVID-19 Open Research Dataset" (CORD-19), which includes over 50,000 articles [46]. Lopez et al. employ the "PubMed Central (PMC) Open Access Subset" dataset for medical software citations [43], while Schindler et al. present the "SoMeSci" benchmark dataset for software mentions in medical articles. While these efforts demonstrate the potential for entity linking in healthcare, broader applications, such as linking clinical trial data or pharmaceutical patents, remain underexplored.

In summary, the available datasets predominantly focus on news and questions, aligning with the primary domains of existing entity linking approaches. While this indicates no significant domain

[30] http://lcg-www.uia.ac.be/conll2003/ner

gap between datasets and approaches, expanding into additional domains, such as mathematics or programming code, could significantly enhance the scope and impact of entity linking research.

Approaches. A significant research gap in entity linking approaches using Wikidata as the target knowledge base lies in the absence of a structured comparison of the functional overlaps and distinctions among existing systems. This gap hinders the ability to promote reuse and modularity across different systems. To illustrate, several systems partially overlap in their functionality and specialties, yet the specific areas of convergence and divergence are not explicitly examined. For example, Delpeuch introduces *OpenTapioca*, a lightweight and exclusively Wikidata-trainable system that serves as a benchmark for evaluating other entity linking methods [12]. In contrast, the *Falcon* system extends its scope by providing joint entity and relation linking capabilities along with an API for simplified access [66]. While these systems share a focus on leveraging Wikidata, the nature and extent of their overlapping functionalities remain unexplored. Similarly, *Arjun*, proposed by Mulang et al., employs an attentive neural network to enhance performance in noisy and unconventional contexts by incorporating knowledge graph context [52]. This contrasts with *VCG*, presented by Sorokin et al., which targets entity mention detection and disambiguation in question answering and is optimized for diverse entity categories [75]. Despite addressing related challenges, the complementary or redundant aspects of these approaches are not systematically compared. Finally, *KBPearl* offers an end-to-end pipeline for knowledge base population, excelling in augmenting incomplete knowledge bases using vast text corpora [40]. Although this system represents a more comprehensive application, its specific contributions relative to systems like *Arjun* or *VCG* are not clearly delineated. A structured comparison of these systems' functionalities and overlaps would provide the research community with valuable insights into their respective strengths, limitations, and opportunities for reuse. Such meta-analyses could not only facilitate the integration of existing modules but also guide future system development and evaluation.

Datasets. The 11 datasets for entity linking using Wikidata as the target knowledge base predominantly focus on question answering use cases and the news and social media domains, which are areas of significant industrial relevance or commercial interest [27]. While these datasets provide valuable resources, a key research gap lies in the limited diversity of use cases and domains represented, which restricts the applicability of entity linking systems to broader contexts.

For example, Dubey et al. introduce *LC-QuAD*, a dataset designed for question answering, containing 30,000 complex questions, paraphrases, answers, and corresponding SPARQL queries [15]. This dataset is well-suited for evaluating question-answering capabilities but does not extend to other domains such as scientific literature or legal texts.

Similarly, Pontes et al. present *HIPE*, a dataset of historic newspapers designed for identifying historical people, places, and other entities in multilingual historical documents [59]. Although valuable for historical research, its scope is constrained to archival and historical analysis.

The *T-REx* dataset provides large-scale alignments between Wikipedia abstracts and Wikidata triples, comprising 11 million triples aligned with 3.1 million Wikipedia abstracts [17]. While comprehensive, this dataset primarily supports tasks rooted in encyclopedic content, leaving gaps in applicability to real-time or specialized content domains.

The *AIDA CoNLL-YAGO* dataset focuses on robust disambiguation of entities in text, leveraging context from knowledge bases and a novel type of coherence graph [28]. However, it remains centered on text-based disambiguation, lacking representations for multimodal or domain-specific challenges.

Finally, *SimpleQuestions* includes 100k questions aimed at evaluating multitask and transfer learning for simple question answering under large-scale conditions [4]. This dataset is instrumental

for benchmarking question-answering models but does not address more complex or nuanced entity linking tasks in niche domains.

The research community would benefit from the development of datasets that cater to a broader range of use cases and domains. Expanding beyond question answering, news, and social media to include areas such as healthcare, scientific literature, and legal texts would significantly enhance the versatility and impact of entity linking systems.

Applications. The most frequent applications of entity linking identified in the reviewed literature include *knowledge base population or knowledge graph population*, *question answering over knowledge graphs*, *relation extraction*, *stance detection*, and *entity extraction*. While question answering and knowledge base population are well-established use cases, other applications, such as stance detection, remain underexplored despite their industrial relevance. An exemplary approach for *knowledge base population or knowledge graph population* is presented by Lin et al., who use joint entity and relation linking, including side information from documents, to populate a knowledge base [40]. Benchmarks for knowledge base population include the TAC-KBP knowledge base population [17] and the "KnowledgeNet" dataset [11]. González et al. populate skill ontologies [21], while Luggen et al. assess knowledge base population completeness using statistical class cardinality calculations [44]. Other works populate narratives using Wikidata events [45] or build medical knowledge graphs to combat the COVID-19 pandemic [60]. Despite the diversity of these efforts, more work is needed to standardize and evaluate knowledge base population methods across domains. *Question answering over knowledge graphs* is prominently addressed by systems such as Falcon 2.0, which jointly links entities and relations in short text questions [66]. Sorokin and Gurevych propose a neural architecture for combined entity mention detection and disambiguation [75], while Liu et al. evaluate entity linking in scenarios where question and answer entities are semantically related but not identical [42]. These approaches demonstrate progress in handling structured question answering but leave conversational and multi-turn question answering largely unaddressed. *Relation extraction* is tackled by Sakor et al., who transform short texts into ranked lists of candidate entities and relations [66]. The end-to-end entity linking system "KBPearl" retrieves canonicalized triples without prior specification of predicates of interest [40]. Similarly, the "RLSW" introduced by Yang et al. clusters relation mentions using a phrase similarity algorithm [87]. However, existing methods primarily focus on short texts, with limited exploration of relation extraction in longer or more complex documents. *Stance detection* is discussed by Hamdi et al., who present a multilingual dataset of Open Government Data annotated for stance detection as a sequence pair classification task [26]. While promising, stance detection applications remain limited in scope and warrant further research to address emerging use cases such as detecting sentiment or bias in diverse media formats. *Entity extraction* for historical documents, such as newspapers and letters in multiple European languages, is discussed by Pontes et al. They highlight challenges posed by language variations and OCR errors [58]. These challenges underscore the need for robust, domain-specific methods to improve performance in underrepresented contexts.

In conclusion, question answering remains a dominant use case for entity linking, alongside knowledge base population, which can, in turn, enhance question answering systems. Other applications, such as stance detection, have significant industrial importance but are still underexplored. Expanding research efforts into these less-studied areas could unlock new opportunities for entity linking systems in both academia and industry.

Challenges. The challenges most frequently discussed in the reviewed literature include *specific approaches*, *knowledge base or knowledge graph evolution*, *datasets*, *ambiguity*, and *sparsity and noise*. While significant progress has been made, several gaps remain that hinder the full potential

of entity linking systems. The challenge of developing *specific approaches* encompasses hybrid solutions combining rule-based and learning-based algorithms [66], entity and relation linking [40], and leveraging advanced methods such as pretrained transformer models (e.g., RoBERTa [61]) and graph algorithms [56]. Despite these advancements, there is still a need for more efficient algorithms capable of handling large-scale datasets with complex interrelations while maintaining computational efficiency. The challenges of *knowledge base or knowledge graph evolution* are particularly pressing. These include instability in links and entity timelines caused by frequent updates [85], targeting multiple knowledge graphs with varying entity formats [52], and dealing with language diversity and entity sparsity [6]. Low-coverage knowledge bases [27], the evolution of historical and geopolitical entities [16], and the Not In Lexicon vulnerability [58] further complicate the task. Addressing these issues requires algorithms that can track and adapt to changes while maintaining historical records for better contextual understanding. *Dataset* challenges primarily involve improving annotation quality and data consistency [21], creating larger and more diverse datasets [27], and addressing metadata quality issues [60]. Without robust benchmarks that include a wider range of use cases, entity linking systems remain limited in their generalizability and domain coverage. The issue of *ambiguity* includes name variation (multiple surface forms for one entity) [71], name ambiguity (one surface form representing multiple entities) [16], and context-dependent meanings [39]. These challenges underscore the importance of context-aware algorithms that can dynamically adapt to different scenarios and contexts. Finally, *sparsity and noise* challenges involve dealing with vocabulary sparsity [16], entity name noise [59], mention heterogeneity [16, 58], digitization noise, and OCR errors or bias [16, 58]. These issues are particularly problematic for historical and low-quality textual data, requiring specialized preprocessing and noise-reduction techniques.

In conclusion, creating more robust datasets and benchmarks that incorporate enhanced (semi-) supervised human quality control is essential. Addressing context disambiguation challenges requires more advanced algorithms, such as **Large Language Models** (**LLMs**), that leverage contextual information from surrounding text, user queries, or historical data. For knowledge graph evolution, agile algorithms capable of tracking changes in entity naming and properties while maintaining historical records of entity timelines would be invaluable. Collaboration among institutions to initiate challenges, develop benchmarks, and expand domain-specific datasets would significantly advance the field of entity linking.

3.3.2 *Comparison to Previous Review.*

Contributions of Previous Review. The previous review of Möller et al. [50] discusses the following research questions: "(1) How do current Entity Linking approaches exploit the specific characteristics of Wikidata? (2) Which unexploited Wikidata characteristics are worth considering for the Entity Linking task? (3) Which Wikidata Entity Linking datasets exist, how widely used are they, and how are they constructed? (4) Do the characteristics of Wikidata matter for the design of Entity Linking datasets and if so, how?". For each reviewed Wikidata entity linking approach, the authors assess the utilized Wikidata characteristics, including labels/aliases, descriptions, knowledge graph structure, hyper-relational structure, entity types, and additional information retrieved from DBpedia or Wikipedia. The authors find that most entity linking approaches do not exploit the specifics of Wikidata, such as its hyper-relational structure, and thus miss out on potential to support their predictions using deep graph information. They argue that Wikidata entity linking could be improved by hyper-relational graph embeddings that include entity-type information or additional textual information retrieved from Wikipedia articles. Moreover, the survey indicates that existing Wikidata entity linking datasets underutilize the platform's capabilities for multilingualism and

time dependence. Specifically, research in this field would greatly benefit from datasets containing documents in multiple languages, as well as dynamic, non-static links. These "transductive" entity embeddings should be designed to evolve alongside the changing Wikidata KG, thus offering a more comprehensive and up-to-date resource for entity linking tasks.

Contribution of this Review. Möller et al. [50] focus on a specific aspect, the exploitation of the hyper-relational structure of Wikidata by entity linking approaches and datasets (two review dimensions). In contrast, we provide an overview of Wikidata entity linking across eight review dimensions. This broader scope allows us to identify research gaps in each of these dimensions, offering a more holistic understanding of the field. We can confirm the central finding of the previous review that classical entity linking approaches do not exploit the hyper-relational structure of Wikidata. The currently available deep learning neural approaches [7, 10, 14, 30, 52, 69, 75] leverage the Wikidata knowledge graph structure with their graph embeddings but do not include multiple layers of hierarchical claims (hyper-relational structure). Furthermore, the challenges we found in the literature and discussed in Sections 3.1.8 and 3.3.1 confirm the time-variance issue also pointed out by Möller et al.. Overall, our research supports the findings of the survey by Möller et al [50]. Moreover, we provide detailed insights into other review dimensions to identify additional research gaps, particularly the need for enhanced *dataset quality* and *hybrid methods* that combine rule-based and learning-based approaches.

3.3.3 Consolidation of Findings. Based on the findings of the previous study and our review, we propose the following list of open tasks for future research on Wikidata entity linking:

— Compile more meta-studies.
— Consolidate definitions of the entity linking tasks or pipeline (canonicalization).
— Foster comparability of methods and their performance.
— Improve approaches by expanding the consideration of descriptions and types of Wikidata articles, the knowledge graph structure and hyper-relational structure of Wikidata, as well as additional textual information from Wikidata articles.
— Enhance the use of fine-grained types, such as software or music.
— Address additional domains, e.g., mathematics, physics, chemistry, or programming code.
— Increase research on less frequently explored but economically relevant applications, such as stance detection.
— Tackle challenges, such as knowledge graph evolution and link instability, dataset (annotation) quality, multilingualism (approaches and datasets) as well as entity sparsity and noise.

We hope that researchers in the field of Wikidata entity linking will consider and address the identified gaps. Particular emphasis should be placed on improving the comparability of entity linking definitions and approaches, knowledge graph evolution, dataset quality, and multilingualism, as these aspects critically affect the performance of entity linking approaches.

3.4 Answers to Research Questions

Based on the results of the Research Tasks (RT 1 - RT 3) in the previous sections, we can answer our Research Questions as follows.

(1) **What do researchers need to know about entity linking with Wikidata in terms of definitions, tasks, types, domains, approaches, datasets, and applications?** *The reviewed literature reveals inconsistent definitions of entity linking, with 37 variations and frequent interchange of terms, such as recognition, identification, and detection, as well as surface forms, mentions, and labels. Despite this variability, most definitions agree that entity linking is*

an end-to-end process involving multiple steps; 36 task descriptions highlight common subtasks, such as recognition, disambiguation, and linking, though these are often merged, renamed, or skipped. The review identified 82 entity types, revealing a mismatch between those emphasized in research, such as humans, and their actual representation in Wikidata, where such types are relatively scarce. Additionally, 44 domains and 34 approaches demonstrate the conceptual and methodological breadth of the field, while 17 datasets and 23 applications underscore the growing importance of robust resources and practical implementations. Together, these findings highlight the need for greater standardization and alignment across definitions, tasks, types, and datasets to enhance the effectiveness and applicability of entity linking research.

(2) **Where is the research need and potential (research gaps)?** *Future research on Wikidata entity linking should focus on standardizing task definitions, improving method comparability, and enhancing techniques through deeper integration of Wikidata's structure and content. It should also expand into underexplored domains and applications, refine the use of fine-grained types, and address key challenges, such as multilingualism, knowledge graph evolution, and data quality. Meta-analyses and studies on economically relevant but overlooked tasks are also encouraged.*

3.5 Limitations

The quality of sources in entity linking with Wikidata varies considerably, as the considered publications appear in journals, conferences, or workshops, which differ in review rigor. There is also a strong bias toward English-language datasets and certain domains, which limits the generalizability of results across languages and application areas. Additionally, positive results are more likely to be published, introducing potential publication bias. Many studies evaluate on static benchmark datasets, which do not fully capture the dynamic and evolving nature of Wikidata or real-world text environments. As a result, reported performance may not transfer reliably to practical applications involving noisy or multilingual data.

4 Outlook

We conclude our review by summarizing our results and offering a brief outlook on potential and planned future work in the field.

4.1 Conclusion

Our review shows that entity linking is most commonly defined as a suite of techniques designed to align unstructured texts with structured concept representations in a knowledge base or knowledge graph. This alignment facilitates information extraction and enables specific queries, such as those from question answering systems. Key tasks within the entity linking domain include the end-to-end pipeline, which encompasses entity mention span detection, type classification, candidate generation, ranking and selection, entity disambiguation, linking to the knowledge base, and the detection of unlikable entities that are labeled as "Not in Lexicon". Moreover, a variety of open source approaches and systems are available. They perform, e.g., simple and fast synchronous Wikidata entity recognition ("OpenTapioca") [12], joint entity and relation linking ("Falcon") [66], and extraction from noisy data to distill canonicalized facts ("KBPearl") [40]. In addition, we find 11 public Wikidata entity linking datasets containing, e.g., complex questions ("LC-QuAD") [15], historical newspapers ("HIPE") [16], or knowledge triples ("T-REx") [17]. Most entity linking approaches focus on the primary entity types: person (PER), organization (ORG), and location (LOC). However, there are also less common fine-grained types, such as songs or software. The domains most frequently addressed include news and tweets, questions,

articles, and medical texts. The applications discussed span across knowledge base population or knowledge graph population, knowledge graph question answering, relation extraction, and stance detection. The challenges identified in these areas mainly include the lack of specific approaches, issues related to knowledge bases or graphs, dataset quality, ambiguity, sparsity, and noise.

4.2 Future Work

We plan to compile a follow-up literature review in due time to examine and illustrate the evolution of Wikidata entity linking research and determine which of the identified research gaps have been addressed. To facilitate this task, we plan to employ automated review generation methods that utilize literature search engine APIs combined with the summarizing capabilities of LLMs, such as the rapidly advancing **General Pretrained Transformer** (**GPT**) models [55]. This integration of technology and research will provide a more dynamic and comprehensive overview of the progress in the field of Wikidata entity linking. We believe that the trend towards generating more frequent or even live updates of reviews will significantly aid researchers in staying up-to-date with recent developments in their fields. This approach will be particularly beneficial for undergraduate and doctoral students, helping them avoid expending resources on redundant research due to the absence of a profound, comprehensive, and current state-of-the-art review of their research topic. We anticipate that these improvements will render research more effective, comparable, and expedited, contributing positively to the research community and enhancing the quality and relevance of scholarly work.

References

[1] Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L. Opdahl, and Csaba Veres. 2020. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access* 8 (2020), 32862–32881. DOI: https://doi.org/10.1109/ACCESS.2020.2973928

[2] Artem Alekseev, Mikhail Chaichuk, Miron Butko, Alexander Panchenko, Elena Tutubalina, and Oleg Somov. 2025. The benefits of query-based KGQA systems for complex and temporal questions in LLM era. In *NLDB (1) (Lecture Notes in Computer Science, Vol. 15836)*. Springer, 426–441.

[3] Debayan Banerjee, Debanjan Chaudhuri, Mohnish Dubey, and Jens Lehmann. 2020. PNEL: Pointer network based end-to-end entity linking over knowledge graphs. In *ISWC (1) (Lecture Notes in Computer Science, Vol. 12506)*. Springer, 21–38.

[4] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. arXiv:1506.02075. Retrieved from https://arxiv.org/abs1506.0207

[5] Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José G. Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust named entity recognition and linking on historical multilingual documents. In *CLEF (Working Notes) (CEUR Workshop Proceedings, Vol. 2696)*. CEUR-WS.org.

[6] Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity linking in 100 languages. In *EMNLP (1)*. Association for Computational Linguistics, 7833–7845.

[7] Lorenzo Canale, Pasquale Lisena, and Raphaël Troncy. 2018. A novel ensemble method for named entity recognition and disambiguation based on neural network. In *ISWC (1) (Lecture Notes in Computer Science, Vol. 11136)*. Springer, 91–107.

[8] Emmanuel Cartier and Emile Peetermans. 2024. Combining deep learning models and lexical linked data: Some insights from the development of a multilingual news named entity recognition and linking dataset. In *Proceedings of the Workshop on Deep Learning and Linked Data (DLnLD)@ LREC-COLING 2024*. 31–44.

[9] Alberto Cetoli, Mohammad Akbari, Stefano Bragaglia, Andrew D. O'Harney, and Marc Sloan. 2018. Named entity disambiguation using deep learning on graphs. arXiv:1810.09164. Retrieved from https://arxiv.org/abs1810.09164

[10] Alberto Cetoli, Stefano Bragaglia, Andrew D. O'Harney, Marc Sloan, and Mohammad Akbari. 2019. A neural approach to entity linking on wikidata. In *ECIR (2) (Lecture Notes in Computer Science, Vol. 11438)*. Springer, 78–86.

[11] Filipe de Sá Mesquita, Matteo Cannaviccio, Jordan Schmidek, Paramita Mirza, and Denilson Barbosa. 2019. KnowledgeNet: A benchmark dataset for knowledge base population. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 749–758.

[12] Antonin Delpeuch. 2020. OpenTapioca: Lightweight entity linking for wikidata. In *Wikidata@ISWC (CEUR Workshop Proceedings, Vol. 2773)*. CEUR-WS.org.
[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.
[14] Daniel Alejandro Diomedi Pinto. 2021. Question answering over wikidata using entity linking and neural semantic parsing. Master's Thesis. University of Chile.
[15] Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A large dataset for complex question answering over wikidata and DBpedia. In *ISWC (2) (Lecture Notes in Computer Science, Vol. 11779)*. Springer, 69–78.
[16] Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers. In *CLEF (Working Notes) (CEUR Workshop Proceedings, Vol. 2696)*. CEUR-WS.org.
[17] Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *LREC*. European Language Resources Association (ELRA).
[18] Nicholas FitzGerald, Daniel M. Bikel, Jan A. Botha, Daniel Gillick, Tom Kwiatkowski, and Andrew McCallum. 2021. MOLEMAN: Mention-only linking of entities with a mention annotation network. In *ACL/IJCNLP (2)*. Association for Computational Linguistics, 278–285.
[19] Johanna Geiß and Michael Gertz. 2016. With a little help from my neighbors: Person name linking using the wikipedia social network. In *WWW (Companion Volume)*. ACM, 985–990.
[20] Johanna Geiß, Andreas Spitz, and Michael Gertz. 2017. NECKAr: A named entity classifier for wikidata. In *GSCL (Lecture Notes in Computer Science, Vol. 10713)*. Springer, 115–129.
[21] Lino González, Elena García Barriocanal, and Miguel-Ángel Sicilia. 2020. Entity linking as a population mechanism for skill ontologies: Evaluating the use of ESCO and wikidata. In *MTSR (Communications in Computer and Information Science, Vol. 1355)*. Springer, 116–122.
[22] Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: A systematic review. *Comput. Sci. Rev.* 29 (August 2018), 21–43.
[23] Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*. 466–471. Retrieved from https://aclanthology.org/C96-1079/
[24] Kensho R&D group. 2020. Kensho Derived Wikimedia Dataset. Retrieved from https://www.kaggle.com/kenshoresearch/kensho-derived-wikimedia-data. Accessed on April 4, 2023.
[25] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with wikipedia. *Artif. Intell.* 194 (2013), 130–150. DOI: https://doi.org/10.1016/j.artint.2012.04.005
[26] Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, José G. Moreno, and Antoine Doucet. 2021. A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *SIGIR*. ACM, 2328–2334.
[27] Bahareh Harandizadeh and Sameer Singh. 2020. Tweeki: Linking named entities on twitter to a knowledge graph. In *W-NUT@EMNLP*. Association for Computational Linguistics, 222–231.
[28] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *EMNLP*. ACL, 782–792.
[29] Binxuan Huang, Han Wang, Tong Wang, Yue Liu, and Yang Liu. 2020. Entity linking for short text using structured knowledge graph via multi-grained text matching. In *INTERSPEECH*. ISCA, 4178–4182.
[30] Filip Ilievski, Eduard H. Hovy, Piek Vossen, Stefan Schlobach, and Qizhe Xie. 2020. The role of knowledge in determining identity of long-tail entities. *J. Web Semant.* 61-62 (2020), 100565.
[31] Anastasiia Iurshina, Jiaxin Pan, Rafika Boutalbi, and Steffen Staab. 2022. NILK: Entity linking dataset targeting NIL-linking cases. In *CIKM*. ACM, 4069–4073.
[32] Lucie-Aimée Kaffee, Hady ElSahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon S. Hare, and Elena Simperl. 2018. Mind the (language) gap: Generation of multilingual wikipedia summaries from wikidata for articleplaceholders. In *ESWC (Lecture Notes in Computer Science, Vol. 10843)*. Springer, 319–334.
[33] Barbara A. Kitchenham and S. Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report EBSE-2007-01. School of Computer Science and Mathematics, Keele University, Keele, UK. Retrieved from https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf. Accessed 2025-07-17.

[34] Marcus Klang and Pierre Nugues. 2020. Hedwig: A named entity linker. In *LREC*. European Language Resources Association, 4501–4508.
[35] Anders Kofod-Petersen. 2018. *How to do a Structured Literature Review in Computer Science*. Technical Report, Version 0.2. Norwegian University of Science and Technology, Dep. of Computer and Information Science, IDI, Trondheim, Norway. Retrieved from https://research.idi.ntnu.no/aimasters/files/SLR_HowTo2018.pdf. Accessed 2025-07-17.
[36] Kai Labusch and Clemens Neudecker. 2020. Named entity disambiguation and linking historic newspaper OCR with BERT. In *CLEF (Working Notes) (CEUR Workshop Proceedings, Vol. 2696)*. CEUR-WS.org.
[37] Tuan Manh Lai, Heng Ji, and ChengXiang Zhai. 2022. Improving candidate retrieval with entity profile generation for wikidata entity linking. In *ACL (Findings)*. Association for Computational Linguistics, 3696–3711.
[38] Xueling Lin and Lei Chen. 2019. Canonicalization of open knowledge bases with side information from the source text. In *ICDE*. IEEE, 950–961.
[39] Xueling Lin, Lei Chen, and Chaorui Zhang. 2021. TENET: Joint entity and relation linking with coherence relaxation. In *SIGMOD Conference*. ACM, 1142–1155.
[40] Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. 2020. KBPearl: A knowledge base population system supported by joint entity and relation linking. *Proc. VLDB Endow.* 13, 7 (2020), 1035–1049.
[41] Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. *Trans. Assoc. Comput. Linguistics* 3 (2015), 315–328. DOI: https://doi.org/10.1162/tacl_a_00141
[42] Cao Liu, Shizhu He, Hang Yang, Kang Liu, and Jun Zhao. 2017. Unsupervised joint entity linking over question answering pair with global knowledge. In *CCL (Lecture Notes in Computer Science, Vol. 10565)*. Springer, 273–286.
[43] Patrice Lopez, Caifan Du, Johanna Cohoon, Karthik Ram, and James Howison. 2021. Mining software entities in scientific literature: Document-level NER for an extremely imbalance and large-scale task. In *CIKM*. ACM, 3986–3995.
[44] Michael Luggen, Djellel Eddine Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. 2019. Non-parametric class completeness estimators for collaborative knowledge graphs—the case of wikidata. In *ISWC (1) (Lecture Notes in Computer Science, Vol. 11778)*. Springer, 453–469.
[45] Daniele Metilli, Valentina Bartalesi, Carlo Meghini, and Nicola Aloia. 2019. Populating narratives using wikidata events: An initial experiment. In *IRCDL (Communications in Computer and Information Science, Vol. 988)*. Springer, 159–166.
[46] Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, Tobias Mayer, et al. 2020. Covid-on-the-web: Knowledge graph and services to advance COVID-19 research. In *ISWC (2) (Lecture Notes in Computer Science, Vol. 12507)*. Springer, 294–310.
[47] Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, Mário J. Silva, Alberto H. F. Laender, Ricardo A. Baeza-Yates, Deborah L. McGuinness, Bjørn Olstad, Øystein Haug Olsen, and André O. Falcão (Eds.). ACM, 233–242. DOI: https://doi.org/10.1145/1321440.1321475
[48] Zhaoyan Ming and Tat-Seng Chua. 2015. Resolving polysemy and pseudonymity in entity linking with comprehensive name and context modeling. *Inf. Sci.* 307 (2015), 18–38. DOI: https://doi.org/10.1016/j.ins.2015.02.025
[49] Fumiya Mitsuji, Sudesna Chakraborty, Takeshi Morita, Shusaku Egami, Takanori Ugai, and Ken Fukuda. 2024. Entity linking for wikidata using large language models and wikipedia links. In *CANDARW*. IEEE, 144–149.
[50] Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2022. Survey on english entity linking on wikidata: Datasets and approaches. *Semantic Web* 13, 6 (2022), 925–966.
[51] Isaiah Onando Mulang', Kuldeep Singh, Chaitali Prabhu, Abhishek Nadgeri, Johannes Hoffart, and Jens Lehmann. 2020. Evaluating the impact of knowledge graph context on entity disambiguation models. In *CIKM*. ACM, 2157–2160.
[52] Isaiah Onando Mulang', Kuldeep Singh, Akhilesh Vyas, Saeedeh Shekarpour, Maria-Esther Vidal, and Sören Auer. 2020. Encoding knowledge graph entity aliases in attentive neural network for wikidata entity linking. In *WISE (1) (Lecture Notes in Computer Science, Vol. 12342)*. Springer, 328–342.
[53] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.
[54] Kristian Noullet, Rico Mix, and Michael Färber. 2020. KORE 50^{DYWC}: An evaluation data set for entity linking based on DBpedia, YAGO, Wikidata, and Crunchbase. In *LREC*. European Language Resources Association, 2389–2395.
[55] OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774
[56] Drew Perkins. 2020. Separating the Signal from the Noise: Predicting the Correct Entities in Named-Entity Linking.
[57] Katherine Louise Polley, Vivian Tompkins, Brendan John Honick, and Jian Qin. 2021. Named entity disambiguation for archival collections: Metadata, Wikidata, and Linked data. In *ASIST (Proc. Assoc. Inf. Sci. Technol., Vol. 58)*. Wiley, 520–524.
[58] Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, José G. Moreno, Emanuela Boros, Ahmed Hamdi, Antoine Doucet, Nicolas Sidere, and Mickaël Coustaty. 2022. MELHISSA: A multilingual entity linking architecture for historical press articles. *Int. J. Digit. Libr.* 23, 2 (2022), 133–160.

[59] Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, José G. Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. 2020. Entity linking for historical documents: Challenges and solutions. In *ICADL (Lecture Notes in Computer Science, Vol. 12504)*. Springer, 215–231.

[60] Jan Portisch, Omaima Fallatah, Sebastian Neumaier, Mohamad Yaser Jaradeh, and Axel Polleres. 2020. Challenges of linking organizational information in open government data to knowledge graphs. In *EKAW (Lecture Notes in Computer Science, Vol. 12387)*. Springer, 271–286.

[61] Vera Provatorova, Svitlana Vakulenko, Evangelos Kanoulas, Koen Dercksen, and Johannes M. van Hulst. 2020. Named entity recognition and linking on historical newspapers: UvA.ILPS & REL at CLEF HIPE 2020. In *CLEF (Working Notes) (CEUR Workshop Proceedings, Vol. 2696)*. CEUR-WS.org.

[62] Henry Rosales-Méndez, Aidan Hogan, and Barbara Poblete. 2020. Fine-grained entity linking. *J. Web Semant.* 65 (2020), 100600. DOI: https://doi.org/10.1016/j.websem.2020.100600

[63] Marco Rospocher and Francesco Corcoglioniti. 2020. Knowledge-driven joint posterior revision of named entity classification and linking. *J. Web Semant.* 65 (2020), 100617.

[64] Adam Aron Rynkiewicz, Raúl Palma, and Piotr Formanowicz. 2025. Universal entity linking. *Eng. Appl. Artif. Intell.* 161 (2025), 112185.

[65] Adam Aron Rynkiewicz, Raúl Palma, and Paulina Poniatowska-Rynkiewicz. 2025. Introducing FELA - flexible entity linking approach. In *HybridAIMS+CAI (Selected Papers) (CEUR Workshop Proceedings, Vol. 3996)*. CEUR-WS.org, 3–10.

[66] Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. 2020. Falcon 2.0: An entity and relation linking tool over wikidata. In *CIKM*. ACM, 3141–3148.

[67] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *CoNLL*. ACL, 142–147.

[68] David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 2021. SoMeSci- A 5 star open data gold standard knowledge graph of software mentions in scientific articles. In *CIKM*. ACM, 4574–4583.

[69] Özge Sevgili, Artem Shelmanov, Mikhail Y. Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web* 13, 3 (2022), 527–570.

[70] Abdul Lathif Fathima Shanaz and Roshan G. Ragel. 2021. Wikidata based person entity linking in news articles. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, 66–70.

[71] Fathima Shanaz and Roshan G. Ragel. 2020. Wikidata based location entity linking. In *ICSCA*. ACM, 307–312.

[72] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* 27, 2 (2015), 443–460.

[73] Anastasia Shimorina, Johannes Heinecke, and Frédéric Herledan. 2022. Knowledge extraction from texts based on wikidata. In *NAACL-HLT (Industry Papers)*. Association for Computational Linguistics, 297–304.

[74] Daniil Sorokin and Iryna Gurevych. 2017. End-to-end representation learning for question answering with weak supervision. In *SemWebEval@ESWC (Communications in Computer and Information Science, Vol. 769)*. Springer, 70–83.

[75] Daniil Sorokin and Iryna Gurevych. 2018. Mixing context granularities for improved entity linking on question answering data across entity categories. In **SEM@NAACL-HLT*. Association for Computational Linguistics, 65–75.

[76] Andreas Spitz, Satya Almasian, and Michael Gertz. 2017. EVELIN: Exploration of event and entity links in implicit networks. In *WWW (Companion Volume)*. ACM, 273–277.

[77] Andreas Spitz and Michael Gertz. 2018. Entity-centric topic extraction and exploration: A network-based approach. In *ECIR (Lecture Notes in Computer Science, Vol. 10772)*. Springer, 3–15.

[78] Yu Su, Huan Sun, Brian M. Sadler, Mudhakar Srivatsa, Izzeddin Gur, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for QA evaluation. In *EMNLP*. The Association for Computational Linguistics, 562–572.

[79] Thang Hoang Ta and Chutiporn Anutariya. 2014. A model for enriching multilingual wikipedias using infobox and wikidata property alignment. In *JIST (Lecture Notes in Computer Science, Vol. 8943)*. Springer, 335–350.

[80] Chuanqi Tan, Furu Wei, Pengjie Ren, Weifeng Lv, and Ming Zhou. 2017. Entity linking for queries by searching wikipedia sentences. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 68–77. DOI: https://doi.org/10.18653/v1/d17-1007

[81] Thomas Pellissier Tanon, Denny Vrandecic, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From freebase to wikidata: The great migration. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*. Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 1419–1428. DOI: https://doi.org/10.1145/2872427.2874809

[82] Nicolas Tempelmeier and Elena Demidova. 2021. Linking OpenStreetMap with knowledge graphs - Link discovery for schema-agnostic volunteered geographic information. *Future Gener. Comput. Syst.* 116 (2021), 349–364.

[83] Ruben Van Heusden, Maarten Marx, and Jaap Kamps. 2022. Entity linking in the ParlaMint corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*. 47–55.

[84] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. DOI: https://doi.org/10.1145/2629489
[85] Albert Weichselbraun, Philipp Kuntschik, and Adrian M. P. Brasoveanu. 2018. Mining and leveraging background knowledge for improving named entity linking. In *WIMS*. ACM, 27:1–27:11.
[86] Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *COLING*. Association for Computational Linguistics, 2145–2158.
[87] Xi Yang, Shiya Ren, Yuan Li, Ke Shen, Zhixing Li, and Guoyin Wang. 2017. Relation linking for wikidata using bag of distribution representation. In *NLPCC (Lecture Notes in Computer Science, Vol. 10619)*. Springer, 652–661.
[88] Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *ACL (2)*. The Association for Computer Linguistics.
[89] Xingchen Zhou, Peng Wang, Guozheng Li, Jiafeng Xie, and Jiangheng Wu. 2021. Weibo-MEL, wikidata-MEL and Richpedia-MEL: Multimodal entity linking benchmark datasets. In *CCKS (Communications in Computer and Information Science, Vol. 1466)*. Springer, 315–320.

Appendix

A Appendix

This Appendix describes the details of our evaluation.

Abbreviations and Publications. Table 7 alphabetically lists the abbreviations we use in this article. In the appendix, we substitute numerical citation markers, which denote publication positions in the reference list, with numerical IDs. This approach facilitates the aggregation of publications into clusters characterized by ascending IDs. Table 8 shows the mapping of citation markers to publication IDs. Table 9 contains type classifications of the publications considered in this review. The types are sorted by their occurrence frequency. Our review covers 33 publications that present approaches, 17 publications that provide datasets, four studies, and one thesis.

Overview of Results. Table 10 shows an overview of the results as a numeric frequency count across the entity linking dimensions of the publications considered in this review. For the full literal table, see https://zenodo.org/records/17839768.

Definitions, Tasks, Subtasks. Tables 11, 12, and 13 respectively show exemplary definitions, tasks, and subtasks extracted from the publications considered in this review.

Approaches and Datasets. Table 14 shows the considered publication types (approach, dataset, study, thesis), and the approaches and/or datasets (if any) the publications present. Some approaches have proper names; to others, we refer by means of the authors' names. Several publications focus on the introduction of approaches, others on datasets, some on both.

Types and Domains. Table 15 provides an overview of the most frequently occurring entity types in the reviewed publications. Table 16 shows a ranking of the extracted entity types included in the reviewed publications. Table 17 is a consolidated view of the entity types after mapping, e.g., fictional character to person, company to organization, and city to location. Table 19 contains the domain statistics of the reviewed Wikidata entity linking publications. News, being by far the most frequent domain, is followed by other articles, such as research and Wikipedia. At the bottom end of the distribution, there is a large number of domain categories that only appear in one publication, e.g., mathematics or physics. Table 18 shows domains (and applications) per publication ID.

Applications. Table 20 shows entity linking applications in decreasing order of the frequency with which we identified them in the reviewed literature. The foremost applications, named entity recognition and named entity disambiguation, significantly overlap with entity linking. The most common "external' applications identified are knowledge graph population and question answering. Table 21 shows the entity linking applications together with the publications, in which they are discussed.

Challenges. Table 22 displays (shortened) descriptions of challenges related to entity linking, sorted by their occurrence frequency in the reviewed publications. Table 23 shows (shortened) examples of challenges together with the number of publications, in which they occur.

Table 7. All **Abbreviations** Used in This Review, Sorted Alphabetically

Application	Abbreviation
Cross-Lingual Named Entity Linking	XEL
Entity Disambiguation	ED
Entity Embeddings	EE
Entity Resolution	ER
Knowledge Base Population	KBP
Knowledge Graph Population	KGP
Multilingual Entity Linking	MEL
Named Entity Disambiguation	NED
Named Entity Linking	NEL
Named Entity Recognition	NER
Named Entity Recognition and Classification	NERC
Natural Language Generation	NLG
Natural Language Processing	NLP
Question Answering	QA
Question Answering over Knowledge Graphs	KGQA
Relation Classification	RC
Relation Extraction	RE
Relation Linking	RL
Stance Detection	StD

Table 8. Mapping Publication IDs to Publication Citations

Publication IDs	Publication Citations
1, 2, 3, 4, 5, 6, 7, 8, 9, 10	[3, 9, 12, 29, 40, 51, 52, 56, 66, 75]
11, 12, 13, 14, 15, 16, 17, 18, 19, 20	[5], [61], [36], [6], [34], [27], [17], [38], [12], [15]
21, 22, 23, 24, 25, 26, 27, 28, 29, 30	[11], [54], [24], [16], [79], [10], [21], [32], [20], [44]
31, 32, 33, 34, 35, 36, 37, 38, 39, 40	[45], [14], [87], [89], [71], [26], [7], [60], [46], [74]
41, 42, 43, 44, 45, 46, 47, 48, 49, 50	[59], [77], [76], [63], [82], [58], [85], [43], [68], [39]
51, 52, 53, 54, 55, 56, 57, 58, 59, 60	[30], [42], [70], [19], [49], [37], [64], [31], [57], [2]
61, 62, 63, 64, 65	[18], [65], [73], [8]

Table 9. Overview of the **Publication Types** Sorted by Occurrence Frequency Rank of the Publications Considered in This Review

Rank	Publication Type	Publication References
1	Approach	[3, 5, 6, 9, 12, 14, 21, 27, 29, 32, 34, 36, 40, 45, 51, 52, 56, 61, 66, 75] [7, 19, 39, 42, 46, 49, 58, 63, 68, 70, 71, 74, 82] [2, 18, 37, 64, 65, 73, 83]
2	Dataset	[10, 11, 15–17, 26, 30, 38, 43, 54, 59, 60, 77, 79, 85]
3	Study	[20, 44, 57, 76, 87]
4	Thesis	[14]

Table 10. Overview of the Results as a Numeric Frequency Count Across the **Entity Linking Dimensions** of the Publications Considered in This Review

Paper ID	Definition(s)	Task(s)	Entity Type(s)	Domain(s)	Approach(es)	Dataset(s)	Application(s)	Challenge(s)
1	1	1	3	2	1	2	2	3
2	2	1	1	1	1	5	5	3
3	1	2	1	1	1	3	3	8
4	1	2	8	1	1	2	2	3
5	1	1	1	2	1	2	2	5
6	2	1	1	1	1	1	1	4
7	2	1	1	1	1	2	2	2
8	3	4	4	1	1	2	2	5
9	2	1	1	1	1	2	2	10
10	1	2	1	1	1	2	2	2
11	2	1	5	1	1	1	1	1
12	1	3	5	1	1	2	2	6
13	3	0	3	1	2	2	2	2
14	1	1	1	1	1	1	1	10
15	1	2	19	5	1	1	1	2
16	1	1	20	1	1	2	2	8
17	0	0	1	1	0	3	3	1
18	1	1	7	5	0	1	1	2
19	1	1	3	1	0	1	1	3
20	0	0	1	1	0	1	1	0
21	0	0	4	2	0	1	1	0
22	0	0	4	1	0	1	1	3
23	0	0	1	1	0	1	1	1
24	1	1	3	1	0	2	2	11
25	0	0	4	1	1	1	1	6
26	1	1	1	1	1	3	3	1
27	0	0	16	1	1	2	2	1
28	0	0	1	1	1	1	1	1
29	1	3	3	1	1	1	1	1
30	0	0	8	1	1	2	2	2
31	0	0	5	1	1	1	1	1
32	1	1	1	1	1	3	3	11
33	1	1	1	1	1	3	3	4
34	1	1	1	3	0	1	1	1
35	1	2	1	1	1	2	2	2
36	2	3	5	1	0	3	3	2
37	1	2	7	2	1	2	2	2
38	0	0	2	1	1	0	1	9
39	1	0	3	1	0	3	3	1
40	0	0	3	1	1	1	1	1
41	1	2	2	1	1	4	4	7
42	0	0	3	1	1	0	3	0
43	1	1	5	1	1	1	1	0
44	0	4	4	1	1	2	2	6
45	1	4	5	1	0	3	3	3
46	1	4	4	1	0	5	5	8
47	1	3	3	2	1	3	3	11
48	1	0	2	2	1	3	3	3
49	0	0	2	9	0	4	4	1
50	0	2	3	9	1	2	2	4
51	1	1	1	1	1	1	1	4
52	1	0	3	1	1	0	3	1
53	1	4	3	1	1	3	3	1
54	0	2	1	1	1	3	3	3
55	1	3	1	1	1	3	2	3
56	1	2	6	3	1	3	3	2
57	1	2	3	3	1	3	0	3
58	1	2	3	1	0	1	1	0
59	1	0	1	1	1	1	0	6
60	1	0	1	1	1	1	0	6
61 - 65	...	...	...	...	...	...	...	...

For the full literal table, see https://zenodo.org/records/17839768.

Table 11. Examples of the Retrieved **Entity Linking Definitions** of the Publications Considered in This Review

Publication ID	Entity Linking Definition
1	Named entity linking is the task of detecting mentions of entities from a knowledge base in free text
2	EL a.k.a. NED is a well-studied research domain for aligning unstructured text to its structured mentions in various knowledge repositories ...
3	EL generally comprises two subtasks: entity recognition that [...], and entity disambiguation that [...]
4	[...] EL is the identification of entity mentions in the question and linking them to entities in KB.
5	[...] linking the noun phrases (resp., relation phrases) detected in the text to the entities (resp., predicates) in the KB.
6	[Linking mentions] to the appropriate entity in the Knowledge Base.
7	[EL consists of two subtasks]: surface form extraction (mention detection) and named entity disambiguation (NED).
8	[EL components]: named-entity recognition, candidate generation, and named-entity disambiguation.
9	[...] linking a mention to the relevant entity is called EL or NED.
10	EL is the task of recognizing named entities in the text and disambiguating them with the corresponding entities in a KG, such as [...].
11 - 65	...

Table 12. Overview of the Available Detected **Entity Linking Task(s)** of the Publications Considered in This Review

Publication ID	Task(s)
1	detecting mentions of KB entities in free text
2	aligning unstructured text to its structured mentions in KBs
3	ER identifying entity surface forms in text, and ED linking the SFs with structures and semi-structured KBs / KGs
4	identification of entity mentions in questions and linking them to entities in KB
5	linking noun phrases to entities in KB
6	link mention (span) to entity in KB / KG
7	link identified (named) entity to ground truth entities in KB
8	identify entities, construct list of candidates, disambiguate, and link to identifier in KG
9	linking mention to relevant entity in KB
10	recognizing (named) entities in text and disambiguating with corresponding entities in KG
11	disambiguation of (named) entities
12	detecting, classifying and linking (named) entities to enable semantic search
14	identify an ungrounded text entity's corresponding entry in a KB
15	automatically finding and linking mentions of things to unique identifiers
16	ground named mentions to a unique entry in KB
18	linking (named) entity mentions in noun phrases
19	detecting mentions of entities from a KB in free text
24	linking entities in text to their corresponding referents in a KB
25	linking an entity mention in some context language to corresponding entity in language-agnostic KB
26	ground named mentions to a unique entry in KB
28	linking mention to relevant entity
31	discover, disambiguate and link surface forms of entity mentions in text to KB
34	linking mentions in text to corresponding entities in KG
35	link surface names in texts to corresponding entity objects in KGs
36	mapping mentions to the corresponding entities in KBs
37	extract mentions in documents, and link them to corresponding entities in a KB
38 - 65	...

Some publications, e.g., Publication ID 13, are missing proper entity linking task descriptions.

Table 13. Overview of the Available Detected **Entity Linking Subtasks** **Definitions** of the Publications Considered in This Review

Publication ID	Subtasks
2	EL = NER + ED
3	EL = ER + ED
6	EL = MD + ED
7	EL = MD + NED
8	EL = NER + CG + NED
13	EL = lookup of possible candidates in index + evaluation of candidates + ranking of candidates
34	EL = candidate entity generation + candidate entity disambiguation and linking
38	NER = NED + EL + StD
39	NERD = NER + NED
45	EL = entity query, sentence query, page query
46	EL = NERD

Table 14. Overview of the Publication Types and Included Approaches and Datasets Considered in This Review

Publication ID	Type	Approach(es)	Dataset(s)
1	Approach	OpenTapioca	RSS-500 news excerpts, ISTEX research article author affiliations
2	Approach	Falcon 2.0	SimpleQuestion, LC-QuAD 2.0, WebQSP-WD
3	Approach	Arjun	T-REx
4	Approach	VCG	WebQuestions, NEEL, GraphQuestion
5	Approach	KBPearl	ReVerb38, NYT2018, LC-QuAD2.0, QUALD-7-Wiki, T-REx, KnowledgeNet, CC-DBP
6	Approach	PNEL	SimpleQuestion, LC-QuAD 2.0
7	Approach	Mulang et al.	Wikidata-Disamb, ISTEX, AIDA-CoNLL
8	Approach	Perkins	KDWD, AIDA CoNLL-YAGO
9	Approach	NED using DL on Graphs	Wiki-Disamb30
10	Approach	Huang et al.	WebQSP
11	Approach	Boros et al.	HIPE
12	Approach	Provatorov et al.	HIPE
13	Approach	Labusch and Neudecker	HIPE
14	Approach	Botha et al.	Mewsli-9
15	Approach	Hedwig	TAC2017
16	Approach	Tweeki	TweekiData, TweekiGold
17	Dataset	-	T-REx
18	Dataset	-	NYT2018
19	Dataset	-	ISTEX-1000
20	Dataset	-	LC-QuAD 2.0
21	Dataset	-	Knowledge Net
22	Dataset	-	KORE50DYWC
23	Dataset	-	Kensho Derived Wikimedia Dataset
24	Dataset	-	CLEF HIPE
25	Dataset	-	Mewsli-9
26	Dataset	-	TweekiData
26	Dataset	-	TweekiGold
27	Approach	Hoang et al.	Infoboxes
28	Approach	Cetoli et al.	JNLPBA, BC2GM, BC5CDR, NCBI-Disease
29	Study	González et al.	ESCO website
30	Study	Kaffee et al.	Kaffee et al.
31	Approach	NECKAr	Wikidata NE
32	Approach	Luggen et al.	Luggen et al.
33	Study	Metilli et al.	WEG
34	Thesis	Pinto et al.	LC-QuAD 2, QALD-7, WikiSPARQL
35	Approach	RLSW	HUMAN
36	Dataset	-	Weibo-MEL, Wikidata-MEL, and Richpedia-MEL
37	Approach	Shanaz & Ragel	AIDA-CoNLL news
38 - 65	...	...	...

Table 15. Overview of the Detected **Entity Types** (Sorted by Publication ID) of the Publications Considered in This Review

Publication ID	Entity Type(s)
1	person, organization, location
2	general
3	general
4	event, location, organization, profession, person, and 3 more
5	general
6	general
7	author-affiliation
8	person, organization, location, miscellaneous
9	general
10	general
11	organization, product, time
12	organization, person, product, location, time
13	person, location, organization
14	general
15	person, organization, company, institution, location, and 14 more
16	person, taxon, film, human settlement, album, and 15 more
17	general
18	person, organization, film, moon, league, airport, location
19	person, organization, location
20	general
21	person, organization, location, date
22	resource, entity, person, organization
23	general
24	organization, product, time
25	general
26	person, taxon, film, human settlement, album, and 15 more
27	species, genus, ordo, familia
28	protein, DNA, RNA, cell line, cell type, gene, disease, chemical
29	company, agency, institution, nationality, religion, political group, and 10 more
30	general
31	person, location, organization
32	video game console, volcano, skyscraper, hospital, mountain, municipality, cathedral, painting
33	work, person, organization, other, event
34	general
35	person
36	person
37	location
38 - 65	...

Table 16. Overview of the Detected **Entity Types (Sorted by Occurrence Frequency Rank)** of the Publications Considered in This Review

Rank	Type	Publication ID(s)
1	person	[1, 4, 8, 11, 12, 13, 15, 16, 18, 19, 21, 22, 26, 29, 31, 33, 35, 36, 38, 39, 43, 44, 45, 46, 48, 49, 52, 53, 54, 55, 56, 57, 58 60, 63, 64, 65]
2	organization	[1, 4, 8, 11, 12, 13, 15, 16, 18, 19, 21, 22, 24, 26, 31, 33, 39, 40, 44, 45, 46, 48, 49, 52, 55, 56, 57, 58, 60, 63, 64, 65]
3	location	[1, 4, 8, 11, 12, 13, 15, 18, 19, 21, 29, 31, 37, 38, 43, 44, 45, 46, 48, 49, 52, 54, 55, 56, 57, 58, 60, 63, 64, 65]
4	product	[4, 11, 12, 24, 29, 48]
5	mountain	[15, 16, 26, 39, 47]
6	time	[11, 12, 24, 38, 60]
7 - 82	...	[...]

Table 17. Consolidated Overview of the Detected **Entity Types** (Sorted by Occurrence Frequency Rank) of the Publications Considered in This Review

Rank	Type	Publication ID(s)
1	person	[1, 4, 8, 11, 12, 13, 15, 16, 18, 19, 21, 22, 26, 29, 31, 33, 35, 36, 38, 39, 41, 43, 44, 45, 46, 48, 49, 52, 53, 54, 55, 56, 57, 58, 60, 63, 64, 65]
2	organization	[1, 4, 7, 8, 11, 12, 13, 15, 16, 18, 19, 21, 22, 24, 26, 29, 31, 33, 38, 39, 40, 44, 42, 45, 46, 47, 48, 49, 52, 55, 56, 57, 58, 60, 63, 64, 65]
3	location	[1, 4, 8, 11, 12, 13, 15, 16, 18, 19, 21, 26, 29, 31, 32, 37, 38, 39, 40, 43, 44, 45, 46, 47, 48, 49, 52, 54, 55, 56, 57, 58, 60, 63, 64, 65]
4	product	[4, 11, 12, 16, 18, 26, 24, 29, 32, 33, 28, 42, 48, 50, 51]
5	time	[4, 11, 12, 16, 21, 24, 26, 29, 33, 38, 45, 60]
6	miscellaneous	[4, 8, 16, 22, 26, 27, 29, 33, 41, 46, 47, 54]
7 - 82	...	[...]

Table 18. Overview of the Detected **Entity Linking Domains and Applications** (Sorted by Publication ID) of the Publications Considered in This Review

Publication ID	Domain(s)	Task(s) / Application(s)
1	news, research articles	ER, EL
2	web questions	ER, EL, RE, RL, QA
3	Wikipedia articles	NER, EL, KGP
4	general	EL, QA
5	news, web questions	RE, KBP
6	web questions	QA
7	general	NED, KGP
8	news	NED, EL
9	general	NED, KGP
10	web questions	EL, KGP
11	historical newspapers	EL
12	historical newspapers	NER, EL
13	Wikipedia articles	NER, EL
14	news	MEL
15	news, discussion forum, web blogs, tweets, research articles	EL EL
16	tweets	NER, EL
17	abstracts	KBP, RE, NLG
18	news, entertainment, business, science, sports	EL
19	research articles	EL
20	general complex questions	QA
21	Wikipedia abstracts, biographical texts	KBP
22	news	EL
23	Wikipedia articles	NLP
24	historical newspapers	ER, EL
25	news in multiple languages	MEL
26	tweets	EL
26	tweets	EL
27	biology	MEL
28	biomedicine	NED
29	skills (ESCO)	EL, KGP
30	open	MEL
31	general	ETC
32	general	EL, KGP
33	events	KGP
34	web questions	EL, QA, KGQA
35	persons	ER, EL, RE, RL, QA
36	social media, encyclopedia, multimodal knowledge graphs	MEL MEL
37	news	EL
38 - 65	...	...

See Table 7 for understanding the application abbreviations.

Table 19. Overview of the Detected **Entity Linking Domains** (Sorted by Occurrence Frequency Rank) of the Publications Considered in This Review

Rank	Domain	Publication ID(s)
1	news	[1, 5, 8, 14, 15, 18, 22, 37, 44, 46, 52, 55, 56, 57, 61, 63, 65]
2	web questions	[2, 5, 6, 10, 34, 54, 60]
3	historical newspapers	[11, 12, 24, 38, 43, 48]
4	research articles	[1, 15, 19, 60]
5	Wikipedia articles	[3, 13, 23]
6	tweets	[15, 16, 26]
7	encyclopedia	[36, 49, 52]
8	medical texts	[41, 50, 51]
9	business	[18, 52]
10	sports	[18, 52]
11 - 44	...	[...]

Table 20. Overview of the Detected **Entity Linking Applications** with Abbreviations, Sorted by the Number of Publications, in Which They are Discussed

Application	Abbreviation	Publications
Knowledge Graph Population	KGP	12
Question Answering	QA	9
Multilingual EL	MEL	7
Relation Extraction	RE	6
Entity Resolution	ER	4
Relation Linking	RL	3
Knowledge Base Population	KBP	3
Entity Type Classification	ETC	3
Question Answering over Knowledge Graphs	KGQA	3
Cross-Lingual Named Entity Linking	XEL	2
Named Entity Linking	NEL	2
Natural Language Generation	NLG	1
Natural Language Processing	NLP	1
Multi-Modal Entity Linking	MMEL	1
Stance Detection	StD	1
Component Extraction	CE	1
Relation Classification	RC	1
Named Entity Recognition and Classification	NERC	1
Entity Embeddings	EE	1
Entity Disambiguation	ED	1
Non-Informative Label IDentification	NILID	1

Table 21. Overview of the Detected Entity Linking **Applications** (Sorted by Occurrence Frequency Rank) of the Publications Considered in This Review

Rank	Application	Publication ID(s)
1	NER	[3, 12, 13, 16, 28, 38, 39, 43, 51, 55, 56]
2	NED	[7, 8, 9, 28, 37, 39, 48, 49, 50, 55, 56]
3	KGP	[3, 7, 9, 10, 29, 32, 33, 40, 41, 47]
4	QA	[2, 4, 6, 20, 34, 54, 56, 58, 60, 64]
5	MEL	[14, 25, 27, 30, 43, 48, 61, 62]
6	RE	[2, 5, 17, 35, 51, 64]
7	ER	[1, 2, 24, 44]
8	RL	[2, 35, 52]
9	KBP	[5, 17, 21]
10	ETC	[31, 44, 47]
11	KGQA	[34, 42, 54]
12	XEL	[43, 48]
13	NEL	[49, 50]
14	NLG	[17]
15	NLP	[23]
16	MMEL	[36]
17	StD	[38]
18	CE	[41]
19	RC	[41]
20	NERC	[46]
21	EE	[48]
22	ED	[51]
23	NILID	[53]

See Table 7 for understanding the application abbreviations.

Table 22. Wikidata **Entity Linking Challenge** Descriptions (Shortened) and Types Sorted Descendingly by the Number of Publications (#) in Which They are Discussed

Nr.	Type	Description (short)	#
1	approaches	Lack of specific approaches needed for special use cases	18
2	knowledge graph	Wikidata KG evolution causes EL problems	17
3	datasets	Data quality issues	13
4	ambiguity	Mapping surface forms to multiple entities	10
5	sparsity	Data sparsity leads to low EL performance	9
6	noise	Problems for EL working on noisy inputs / training data	9
7	quality	Need to control the quality of EL processes	8
8	cross-language	Multi-language EL systems face challenges	8
9	coverage	Difficulties to get full coverage of all potential entities	8
10	generalizability	Poor generalizability to other domains or languages	7
11	preprocessing	Preprocessing tasks, e.g., OCR introduce EL problems	7
12	context	Ignoring context may cause low performance	7
13	language-gap	Focus on English language leads to model bias	7
14	change	EL systems are hard to maintain, extend, and adapt	7
15	annotation	Lack of annotated datasets or wrong annotations	7
16	coherence	High or low coherence between entities	5
17	rareness	Modeling rare entities is difficult	5
18	canonicalization	Challenges for gold standards to be canonicalized	5
19	variation	Name variation and ambiguity challenge EL systems	5
20	augmentation	Difficulties including external information	5
21	encodings	Finding suitable encodings for neural EL is challenging	5
22	vocabulary	Out-of-vocabulary problem	5
23	capitalization	Case sensitivity of surface forms affects EL performance	4
24	benchmarking	Lack of or multiple gold standards for specific EL domains	4
25	absence	Low NIL prediction performance	4
26	length	Long surface forms can be challenging for EL systems	4
27	spelling	Historical documents contain spelling variations	3
28	low-languages	Low-resource languages lead to low EL performance	3
29	granularity	Entity annotation granularity differs with annotators	3
30	wiki-connection	Need to exploit Wikipedia-Wikidata connections more	3
31	comparability	Missing guidelines for EL tasks deteriorate comparability	3
32	implicitness	Sometimes entity mentions are implicit	3
33	cross-referencing	Document-level cross-references	3
34	ranking	Many entity candidates challenge ranking process	2
35	candidate-selection	Difficulities selecting correct entity from candidate lists	2
36	grammar	EL systems need to tackle grammatical mistakes	2
37	pipeline	EL systems not considering task dependencies	2
38	standards	Different EL standards due to missing consensus	2
39	applications	Problems disambiguating entities in applications	1
40	vandalism	Wikidata vandalism harms EL system performance	1
41	typos	Typographical errors in texts cause EL problems	1
42	typification	Different EL types may be confused	1
43 - 65	...	...	1

Table 23. Challenge Examples (Shortened) and Types Sorted Descendingly by the Number of Publications (#), in Which They are Discussed

Nr.	Type	Example (short)	#
1	approaches	Hybrid (rule-based + ML) or long-tail EL systems	18
2	knowledge graph	EL systems rely on the stability and timeliness of wikilinks	17
3	datasets	Data inconsistency (e.g., varying entity names or types)	13
4	ambiguity	'Michael Jordan' basketball or football player	10
5	sparsity	Not enough entity mentions for the EL system to generalize	9
6	noise	Large number of Wikidata entities introduces noise	9
7	quality	Entities extracted from Wikidata	8
8	cross-language	EL templates need to be fine-tuned to new languages	8
9	coverage	New entities or changes in vocabulary challenge EL	8
10	generalizability	Low portability between highly specialized domains	7
11	preprocessing	OCR digitization noise	7
12	context	Without context, linking to the most popular entity	7
13	language-gap	Lower EL system performance in non-English languages	7
14	change	Failing redirects, missing information, or old links	7
15	annotation	'United' instead of 'United Kingdom'	7
16	coherence	Entities in the same context are often topically related	5
17	rareness	Class imbalance for rare entity disambiguation	5
18	canonicalization	Different naming conventions in different languages	5
19	variation	Several surface forms for one entity or vice versa	5
20	augmentation	Exploit external links and references to enrich EL data	5
21	encodings	Less text training data to learn encodings for rare entities	5
22	vocabulary	Rare entities not in knowledge base vocabulary	5
23	capitalization	Organization 'United Nations' (Q1065) vs. 'united nations'	4
24	benchmarking	Linked datasets to use for social media analysis missing	4
25	absence	Missing Wikidata item for 'physics curriculum in Germany'	4
26	length	'Worcester's Breakfast Club for HM Forces and Veterans'	4
27	spelling	Language changes over time (after spelling reforms)	3
28	low-languages	Languages with underserved Wikipedia articles	3
29	granularity	Sometimes only major types, sometimes more finegrained	3
30	wiki-connection	Cross referencing text fragments with Wikipedia pages	3
31	comparability	Each EL system with own taxonomy of entity types	3
32	implicitness	Entities mentioned paraphrased or using pronouns	3
33	cross-referencing	Cross referencing text fragments with Wikipedia pages	3
34	ranking	Considering each word of sentence as entity candidate	2
35	candidate-selection	Correct targets can be below the confidence threshold	2
36	grammar	Orthographic variations introduced by OCR engines	2
37	pipeline	Mention Detection (MD) and Entity Disambiguation (ED)	2
38	standards	Lack of a strict schema and heterogeneity of formats	2
39	applications	QA system misses a significant part of answers	1
40	vandalism	Wikidata items are corrupted or distorted	1
41	typos	'Brian Adams' vs. 'Bryan Adams'	1
42	typification	People names used for street names	1
43-65	...	...	1

Received 9 February 2024; revised 20 December 2025; accepted 22 January 2026

CiteAssist

Citation Sheet

Generated with citeassist.uni-goettingen.de

BibTeX Entry

```
@article{Scharpf2026,
  author={Scharpf, Philipp and Breitinger, Corinna and Spitz,
      Andreas and Meuschke, Norman and Greiner-Petter, André and
      Schubotz, Moritz and Gipp, Bela},
  title={Entity Linking with Wikidata: A Systematic Literature
      Review},
  address={New York, NY, USA},
  journal={ACM Computing Surveys},
  publisher={Association for Computing Machinery},
  topic={nlp},
  year={2026},
  month={01}
}
```