IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 August 2025, accepted 23 August 2025, date of publication 29 August 2025, date of current version 5 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3604258

SURVEY

Multilingual Relation Extraction: A Survey

MANZOOR ALI, RENÉ SPECK, HAMADA M. ZAHERA, MUHAMMAD SALEEM, DIEGO MOUSSALLEM, AND AXEL-CYRILLE NGONGA NGOMO

Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, 33098 Paderborn, Germany

Corresponding author: Manzoor Ali (manzoor@campus.uni-paderborn.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project TRR 318/1 2021–438445824; in part by the German Federal Ministry of Research, Technology and Space (BMFTR) within the Project Künstliche Intelligenz AKADEMIE OWL (KI-AKADEMIE OWL) under Grant 01IS24057B; in part by the EuroStars Project E! 5736 Machine Learning for SPARQL Query Optimization over Centralized and Distributed RDF Knowledge Graphs (SPARQL-ML) under Grant 01QE2429B; and in part by the Open Access Publication Fund of Paderborn University.

ABSTRACT Relation extraction plays a fundamental role in applications of various research fields such as knowledge graph construction, event extraction, and question answering over knowledge graphs, as they often rely on extracting relationships between named entities. Relation extraction has been extensively studied in high-resource languages like English. However, there remains a significant gap in supporting languages with limited resources, defined as those lacking comprehensive annotated corpora, linguistic tools, or pre-trained models, limiting the completeness and accuracy of applications that rely on multilingual data. This paper provides a comprehensive survey of recent advances in relation extraction, focusing on multilingual approaches. We systematically review state-of-the-art methods, datasets used for evaluation, and key features leveraged in these approaches. Additionally, we perform a detailed comparative analysis of the surveyed methods, examining their methodologies, target domains, levels of extraction, explored languages, and effectiveness. Finally, we identify promising directions for future research, with an emphasis on enhancing multilingual relation extraction.

INDEX TERMS Low-resource languages, multilingual, natural language processing, relation extraction, systematic survey.

I. INTRODUCTION

Relation extraction (RE) is a fundamental task within natural language processing (NLP) that aims to discern semantic relationships between entities in text of multiple paragraphs or a single sentence [1], [2], [3], [4]. For example, in ''The European Union's headquarters are situated in Brussels.'' the named entities ''European Union'' and ''Brussels'' are connected by the ''headquarters'' relation. RE holds immense significance for extracting structured information from unstructured textual data, facilitating knowledge graph construction and enhancing natural language understanding systems [5], [6]. RE is pivotal in various domains, including GraphRAG, biomedical research, finance, and customer support automation, aiding in automated processing of vast amounts of text.

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

Despite notable strides in RE techniques, challenges persist due to natural language variability, ambiguity, and contextual dependencies [7]. For instance, extracting the ''headquarters'' relation from the sentence ''The multinational technology conglomerate, Alphabet Inc., which is the parent company of Google, has its parent campus in Mountain View, California, while its European headquarters are in Dublin, Ireland.'' illustrates the complexities involved in accurately identifying such relationships. Specifically, challenges include distinguishing between nested entities (e.g., recognizing ''Alphabet Inc.'' as the parent of ''Google'') and extracting multiple relations, such as the headquarters locations in ''Mountain View, California'' and ''Dublin, Ireland''. The sentence also presents long-range dependencies between entities and requires handling pronoun co-references (e.g., ''its'' referring to ''Alphabet Inc.''), while identifying implied relations like the existence of multiple headquarters further complicates the extraction process. Researchers have explored methodologies

like rule-based approaches, neural networks, and transfer learning to address these challenges [2], [8].

Multilingual RE (MRE) is crucial for delivering comprehensive and nuanced information. By leveraging data from diverse linguistic sources, MRE ensures inclusivity, accuracy, and depth [9]. It enables global reach and local content accessibility, catering to users who do not speak English as their first language. This inclusivity is vital as significant information is often available only in local languages, enhancing the richness of the information pool [10]. Furthermore, MRE captures cultural nuances and cross-verifies information across languages, improving reliability and providing a balanced perspective essential for thorough data analysis.

What fundamentally distinguishes MRE from monolingual RE is its ability to operate across language boundaries, extracting relations from texts in multiple languages using a unified framework. While monolingual RE systems are designed for and optimized on a single language (predominantly English), MRE systems must handle diverse linguistic phenomena, syntactic structures, and semantic expressions that vary significantly across languages. This cross-lingual capability introduces unique challenges not present in monolingual settings:

Linguistic Diversity: Languages differ in word order (Subject-Verb-Object in English vs. Subject-Object-Verb in Japanese), morphological complexity (agglutinative features in Turkish vs. isolating features in Chinese), writing systems (alphabetic, logographic, syllabic), and grammatical features (gender agreement in Romance languages, case marking in Slavic languages) [11].

Resource Disparity and Transfer: High-resource languages like English benefit from abundant labeled data, pre-trained models, and linguistic tools, while low-resource languages often lack these. MRE systems must transfer knowledge across this divide without introducing bias or performance degradation [12], [13].

Cultural and Contextual Variations: Relations may be expressed differently across cultures, requiring systems to understand cultural contexts and nuances that affect relation semantics.

Real-world examples underscore the necessity of MRE. During the COVID-19 pandemic, local language reports provided timely updates and culturally specific advice crucial for effective responses [14]. In disaster management, local language reports offer immediate updates, as seen during the Fukushima nuclear disaster in Japan [15]. In economic contexts, market trends and business intelligence are often published in local languages, such as Chinese financial news, requiring MRE for comprehensive analysis [16]. Political analysis and cultural research also benefit, with local reports providing early insights, as demonstrated by the Arab Spring, predominantly reported in Arabic [17]. These examples highlight that English-only RE systems may not provide a complete picture of global events and trends. The real-world impact of MRE extends beyond these examples, addressing critical limitations of monolingual approaches:

Information Accessibility: MRE democratizes access to knowledge by enabling users to query and retrieve information across language barriers. For example, a researcher can extract biomedical relations from papers published in multiple languages, accessing insights that would otherwise remain siloed.

Comprehensive Knowledge Graphs: Multilingual knowledge graphs constructed using MRE provide a more complete representation of world knowledge by integrating information from diverse linguistic sources. This is particularly valuable for entities and relations that are predominantly discussed in non-English sources.

Cross-cultural Business Intelligence: Companies operating globally can leverage MRE to extract competitive intelligence, consumer sentiment, and market trends from local language sources, gaining insights that would be missed by English-only systems.

Crisis Response: During global crises, MRE enables the rapid extraction and integration of critical information from local reports in multiple languages, facilitating faster and more effective responses.

While considerable headway has been achieved in monolingual RE, transitioning to multilingual settings introduces additional complexities. MRE must address diverse linguistic characteristics, cross-lingual variations in expression, and resource constraints. These challenges persist despite recent advances in multilingual pre-trained language models: [18]

Representation Gaps: Even state-of-the-art multilingual models like XLM-R and mBERT show performance disparities across languages, particularly for typologically distant languages or those underrepresented in pre-training data.

Annotation and Evaluation Challenges: Creating high-quality annotated datasets for multiple languages remains resource-intensive, and cross-lingual annotation transfer often introduces noise. Additionally, the lack of standardized multilingual benchmarks hampers fair evaluation across languages [19].

Domain Adaptation: Adapting MRE systems to specialized domains (e.g., biomedical, legal) across multiple languages presents compounded challenges due to domain-specific terminology and relation types.

Addressing these persistent challenges is essential for developing truly effective MRE systems that can bridge language barriers and provide equitable access to information extraction capabilities globally.

Upon reviewing existing literature, we identified a lack of systematic reviews dedicated to MRE. In the literature, researchers focus on the RE task, information extraction, or combining RE with named entity recognition. Also, different variant of RE (i.e., neural network based approaches, causual RE [2], information extraction as whole [21], open information retrieval based approaches [3]) have extensive surveys available in the literature. Hence, these surveys are

TABLE 1. Overview of peer reviewed surveys on relation extraction. Where OIE (Open Information Extraction), ML/DL (Machine Learning/Deep Learning), DS (Distant Supervision), IE (Information Extraction), MRE (Multilingual Relation Extraction), CRE (Causal Relation Extraction), and TRE (Temporal Relation Extraction) while #D represents the number for multilingual datasets and #A the number of multilingual approaches.

Study	Year	Main Focus	Techniques							#D	#A
			OIE	ML/DL	DS	IE	MRE	CRE	TRE		
This Work	2025	MRE	✓	✓	✓	✗	✓	✓	✓	21	18
Zhao et al. [20]	2024	DL	✓	✓	✗	✗	✓	✓	✗	3	2
Zhou et al. [3]	2022	OIE	✓	✗	✗	✓	✗	✗	✗	0	2
Yang et al. [2]	2022	CRE	✗	✗	✗	✗	✗	✓	✗	0	1
Yang et al. [21]	2022	IE	✗	✓	✗	✓	✗	✗	✗	0	1
Wang et al. [22]	2022	DL	✗	✓	✗	✗	✗	✗	✗	1	1
Nasar et al. [1]	2021	NER & RE	✗	✓	✗	✗	✗	✗	✗	0	0

disjoint from our work, as we target on summarizing existing MRE. Table 1 shows the recent literature studies of RE related approaches and tasks. We rigorously tagged a corpus of papers based on objectives, methodologies, datasets, and evaluation metrics. Despite expanding research, to the best of our knowledge, a survey encompassing state-of-the-art approaches, challenges, and future directions in MRE is missing. Our study aims to fill this void by providing a comprehensive overview of the state-of-the-art approaches for MRE, identify key challenges, and provide novel insights.

TABLE 2. List of abbreviations and descriptions.

Abbreviation	Description
AUC	Area Under the Curve
DL	Deep Learning
EARL	Event Argument Role Labelling
GAN	Generative Adversarial Network
IE	Information Extraction
KG	Knowledge Graph
LLMs	Large Language Models
ML	Machine Learning
MRE	Multilingual Relation Extraction
NER	Named Entity Recognition
NLP	Natural Language Processing
OIE	Open Information Extraction
PLM	Pre-Trained Language Model
RC	Relation Classification
RE	Relation Extraction

Our main contributions can be summarized as follows:

- We provide a **systematic survey** of state-of-the-art approaches defining, understanding, and representing the RE task with a focus on multilingual approaches.
- We **categorized** different MRE approaches according to their main methodology, easing the path for novices.
- We compare the **performance** of selected MRE approaches using metrics such as F1-score, Recall, and Precision.
- We provide a comprehensive comparison of multilingual **datasets and benchmarks** proposed in the literature.
- We highlight **open research questions** and **applications** based on our analysis of MRE.

The rest of this paper is structured as follows: In Section II, we introduce the notation required to understand the paper. In Section III, we present our systematic survey methodology, divided into: 1) related surveys on MRE and 2) publications on MRE approaches. Section IV categorize and analyze the relevant literature and discusses benchmarks as well as datasets. Section V overviews research findings and open questions, and discusses applications. Finally, Section VI concludes and discusses future work.

II. PRELIMINARIES

There are different definitions of RE and related concepts, which originate from various research fields and approaches, e.g., natural language processing (NLP), knowledge graphs (KGs), semantic community, and domain-specific definitions. In this section, we define the terminology and notation used throughout this survey. First, we define general terms, which are common in many approaches. Afterward, we define the terms related to specific categories of approaches.

Table 2 lists symbols used throughout this survey.

A. KNOWLEDGE GRAPH

A KG is a data graph, whose nodes represent entities of interest in the real world and whose edges represent relations between these entities. In general, a KG conforms to a graph-based data model, for example, the RDF model or the property graph model [23].

Let G = (E, R, S) be a KG. E denotes a set of entities (i.e., things of the real world), R denotes a set of relations (i.e., relationships between things), and $S \subseteq E \times R \times E$ denotes a set of triples (i.e., statements about entities and relations), where each triple $(h, r, t) \in S$ contains two entities $h, t \in E$ and a relation $r \in R$.

B. NAMED ENTITY RECOGNITION

Entities are instances of concepts of interest. Named Entity Recognition (NER) is the task of: (i) Identifying mentions of named entities, i.e., proper nouns, in a given text, and (ii) classifying these mentions to predefined categories [24], such as human[1] or location[2] defined by a KG like Wikidata.

[1]https://www.wikidata.org/wiki/Q5
[2]https://www.wikidata.org/wiki/Q17334923

In general [25], given a sentence $\langle w_1, w_2, \ldots, w_n \rangle$ that is a finite sequence of words, i.e, a sentence after tokenization. NER aims to find a sequence $\langle t_1, t_2, \ldots, t_n \rangle$ that holds for each word the predicted category. Consider the following input example: ⟨Albert, nació, en, Ulm⟩. An output of a NER algorithm might be: ⟨human, _, _, location⟩, where Albert denotes that no category was found.

C. ENTITY LINKING AND DISAMBIGUATION

The Entity Linking and Disambiguation (EL) task links each relevant entity mention, for instance Q937, found in a sentence to a descriptor of what that entity mention refers to in the context where it appears [26]. The entity descriptors can be taken, for instance, from a KG.

Given [27] a set of entities $E = \{e_1, e_2, \ldots\}$ (e.g., within a KG) and for each sentence $s = \langle w_1, w_2, \ldots \rangle$ a set of entity mentions $M_s = \{m_1, m_2, \ldots\}$. The entity mentions are given by $m_1 = \langle w_i \rangle_{i=k,\ldots,l}$ and $m_2 = \langle w_j \rangle_{j=m,\ldots,n}$ with $1 \leq k \leq l \leq m \leq n$. The EL task aims to map each entity mention $m \in M_s$ in a sentence s to its corresponding entity $e \in E$, .e.g., Q937 to the Wikidata descriptor nació en.[3]

D. RELATION EXTRACTION AND CLASSIFICATION

The Relation Classification (RC) task identifies and categorizes semantic relations between entities in text. Commonly RC follows a closed setting, i.e., all relations considered as known a priori by a fixed set R, for instance, given by a KG. RC requires pre-learned knowledge, such as surface forms, i.e., relation mentions [28], tree patterns [29], or vector representations [30], which are obtained through RE from text. RE extracts this relevant knowledge, which is then used by RC to classify relationships from text [31]. For instance, with the extracted surface form h —r→ t we might classify the Wikidata relation P19[4] in a sentence in Spanish.

In general [32], [33], given a relation $r \in R$ and a sentence $s = \langle w_1, w_2, \ldots \rangle$ containing at least two entity mentions expressing two entities $e_1, e_2 \in E$, for instance, given by a KG. A mapping function $C(.)$ can be given as

$$C_r(F(s)) = \begin{cases} +1 & \text{if } e_1 \text{ is } r\text{-related to } e_2 \text{ in } s \\ -1 & \text{otherwise} \end{cases}$$

where $F(s)$ are features extracted from s. After performing feature extraction with textual analysis on the given sentence, for instance, POS tagging or dependency parsing, the mapping function $C(.)$ decides if e_1 and e_2 are related to r in s or not. The function $C(.)$ can be constructed as a discriminative classifier by training on a labeled dataset of positive and negative relation examples. RC is a multilingual task if the token sequences come from different languages [33].

[3]https://www.wikidata.org/wiki/Q937
[4]https://www.wikidata.org/wiki/Property:P19

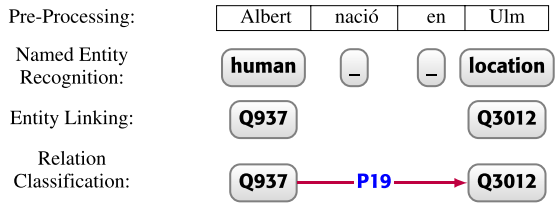


FIGURE 1. An example of four tasks to create a KG triple from text.

Figure 1 shows the example sentence "*Albert nació en Ulm.*" that is processed to create a KG triple[5] from text by a pre-processing step followed by three core extraction tasks: NER, EL, and RC.

Supervised approaches rely on a labeled dataset where relationships between entities are annotated. The model learns to predict relations based on these annotations [32].

Semi-supervised approaches combine labeled and unlabeled data to improve the model's performance. This approach uses a small amount of labeled data and a large amount of unlabeled data [34].

In *distant supervision*, large amounts of labeled data are automatically generated by aligning a large text corpus with a KG, assuming that if a relation exists between two entities in the KG, a sentence mentioning both entities is likely mentioning this relation [35]. Let $B = (e_1, e_2)$ be all sentences in a given corpus mentioning the entities e_1 and e_2, and let $\hat{R} = (e_1, e_2)$ be all relations from e_1 to e_2 in a given KG, i.e., $\hat{R} \subseteq R$. Distant supervision trains with $B = (e_1, e_2)$ and $\hat{R} = (e_1, e_2)$ without sentence level annotations [36].

Unsupervised approaches not rely on labeled data. Instead, it clusters similar contexts to infer relations between entities [37].

E. MULTILINGUAL RELATION EXTRACTION

MRE refers to the process of extracting semantic relationships between entities from texts in multiple languages. This involves handling challenges related to linguistic diversity, such as different grammatical structures, idiomatic expressions, and variations in entity representation across languages.

Given a sentence s with entity mentions from multilingual corpora $\{D_n\}_{n=1}^{N}$, where N represents the number of all languages $\{L_n\}_{n=1}^{N}$. The goal is to extract and predict relations in the same way as monolingual RE but accommodating linguistic variations across languages with methods such as machine translation [38] and prompting PLMs [33], [39].

F. RELATION EXTRACTION LEVELS

Sentence-Level RE involves identifying relationships within a single sentence [40]. *Document-Based* RE extends the task to identify relationships that may be spread across multiple sentences within a document [41]. Consider the following example document D = {"Steve Jobs was a visionary.",

[5]We denote a triple by Steve Jobs, co-founded, Apple

"He co-founded Apple."} with two sentences containing entity mentions and a relation that spans across these sentences. The output of a document-based RE system might be the tuple $T = (\_)$.

G. RELATION EXTRACTION TYPES

General RE aims to extract relations from a wide variety of texts without focusing on a specific domain [42]. In contrast, *Biomedical RE* focuses on extracting relationships between biomedical entities such as genes, proteins, diseases, and drugs from scientific literature [43]. *Business RE* involves identifying relationships relevant to the business domain, such as partnerships, acquisitions, and market analysis [44]. *Temporal RE* involves identifying and classifying temporal relationships between events or entities in a text [45], (e.g., meeting, started after, and lunch). *Hierarchical RE* identifies hierarchical relationships, such as part-whole or subclass relationships. *Causal RE* extracts cause-effect relationships between events or entities [2]. *Conditional RE* identifies conditions under which certain events or relationships hold. *Lexical semantic RE* [46] focuses on identifying fundamental lexical relationships between entities, such as hyponymy (`is-a`), synonymy (`same-as`), antonymy (`opposite-of`), and meronymy (`part-of`) [47].

H. PRE-TRAINED LANGUAGE MODEL

A pre-trained language model (PLM) is a type of neural network model that is trained on a large corpus of text data prior to being fine-tuned on specific NLP tasks. This pre-training helps the model learn general language patterns and representations [48]. A causal PLM is typically trained to minimize the following loss function:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P_\theta(x_t | x_{1:t-1})$$

where x_t is the word at position t in the text sequence.

I. CROSS-LINGUAL TRANSFER LEARNING

Cross-lingual Transfer Learning [49], [50], [51] is a technique where knowledge from a model trained on one language is transferred to help train a model on another language. Min et al. [49] learn discriminative representations to identify semantic relations, regardless of which language the relation mention comes from. Zou et al. [50] transfer feature representations from one language to another. Taghizadeh and Faili [51] utilize representations of sentences that are guaranteed to be consistent across languages.

Mathematically, if θ_L are the parameters of a model trained on language L, these parameters are used as a starting point for training a model on language L'.

For example, BERT trained on English text is a pre-trained model, and using it for fine-tuning on Spanish text is an example of a transfer learning task.

J. MULTILINGUAL EMBEDDINGS

Multilingual Embeddings are representations of words in multiple languages within a shared vector space, enabling the transfer of linguistic knowledge across languages and improving performance on multilingual tasks [52].

K. OPEN INFORMATION EXTRACTION

Open Information Extraction (OIE) extracts n-ary relation tuples from unstructured text, without relying on a predefined ontology schema [3], [53].

Formally, given a sentence as a sequence of words $\langle w_1, w_2, \ldots, w_n \rangle$, OIE outputs a list of tuples $(T_1, T_2, \ldots)$ with the i-th tuple $T_i = (a_{i1}, r_i, a_{i2}, \ldots, a_{iq})$ representing a n-ary relation in the source sequence where r_i denotes the relation in T_i, and a_{ij} is r_i's j-th argument [3]. The arguments are noun phrase found by the OIE system, and the relation is a sequence of words inside the given sentence [53].

Consider the example sentence: *"Deep learning uses multiple layers to extract features from the raw input."*. The extracted tuples by a OIE system might be (Deep learning, uses, multiple layers) and (Deep learning, extracts, features, from the raw input).

L. METRICS

The following metrics have been extensively discussed in the literature for MRE. For a detailed description of these metrics, readers are referred to [54] and [55].

1) PRECISION

Precision measures the proportion of true positive predictions among all positive predictions made by the model:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

2) RECALL

Recall (or Sensitivity) measures the proportion of true positive predictions among all actual positive instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3) MICRO F1 SCORE

The Micro F1 score aggregates contributions from all classes and calculates F1 globally:

$$\text{Micro F1} = 2 \cdot \frac{\text{Micro Precision} \cdot \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}}$$

where Micro Precision and Micro Recall are calculated across all classes:

$$\text{Micro Precision} = \frac{\sum_i \text{TP}_i}{\sum_i (\text{TP}_i + \text{FP}_i)}$$

$$\text{Micro Recall} = \frac{\sum_i \text{TP}_i}{\sum_i (\text{TP}_i + \text{FN}_i)}$$

4) MACRO F1 SCORE

The Macro F1 score calculates F1 for each class independently and then averages them:

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i$$

where F1_i for class i is computed as:

$$\text{F1}_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

5) AREA UNDER THE CURVE (AUC)

AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots True Positive Rate (TPR) against False Positive Rate (FPR) across different thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

where:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Using the one-vs-all strategy, the AUC for each label can be calculated by treating it as a separate binary classification problem and be estimated with [56] and [57]:

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1}$$

where n_0 and n_1 are the numbers of positive and negative examples, respectively, and $S_0 = \sum r_i$, the sum of the ranks of the positive test examples.

$\hat{A}$ is equivalent to the test statistic used in the Mann-Whitney-Wilcoxon two sample test.

III. METHODOLOGY

We follow the structured approaches and guidelines defined by Kitchenham [58] and Moher et al. [59] to conduct our survey study for MRE. In particular, we formulate well-defined research questions, specify our search strategy along with inclusion and exclusion criteria.

A. RESEARCH QUESTIONS

We aim to answer the following research questions (RQs):

RQ1. Which categories of MRE approaches exist, and what are their main building blocks?

RQ2. Which challenges are associated with MRE approaches?

RQ3. What is the (relative) performance of already available MRE approaches?

B. SEARCH STRATEGY

The search strategy employed in this survey was designed to be iterative and independently conducted by the first three authors. This approach mitigates potential biases and ensure comprehensive coverage of relevant literature. In alignment with the predefined research questions, the authors independently selected relevant terms for the review such as "multilingual relation extraction", "multilingual relation finding", "multilingual relation dataset", "low-resource languages relation extraction", "multilingual relation extraction benchmark", and "multilingual information extraction". We used those keywords to derive the following queries:

```
Abstract:
(multilingual relation classification OR
multilingual relation extraction OR
multilingual relationship extraction OR
multilingual information extraction OR
multilingual relation finding) AND
(multilingual approaches OR
multilingual corpora OR
multilingual datasets)
```

The `abstract` tag was used to ensure that the selected keywords in the paper titles are relevant. As suggested [58], [59], a title-based search may not yield a comprehensive collection of relevant studies. Unlike limiting our search to the "intitle" tag, we employ a broader approach for Google Scholar. We refined our search to capture articles or publications that, at a minimum, included the term "multilingual" in their text, aligning the results more closely with our research objectives. Additionally, we noted that many search results were classified under "multilingual information extraction" rather than "multilingual relation extraction" leading to a considerable volume of extraneous results.

C. SEARCH DATABASES

With the predefined keywords in Section III-B, we searched for publications in the following list of search engines, digital libraries, journals, conferences, and their respective workshops:

- Google Scholar[6]
- IEEE Xplore Digital Library (IEEE Xplore)[7]
- ACM DL[8]
- Science Direct[9]
- ACL Anthology (ACL)[10]
- The DBLP Computer Science Bibliography (DBLP)[11]
- Semantic Web Journal (SWJ)[12]

D. INCLUSION/EXCLUSION CRITERIA

This survey focuses on research published between 2018 and 2024 to ensure the inclusion of the most recent advancements in MRE. This time frame captures the significant impact of recent breakthroughs in NLP, particularly those driven by BERT GPT-based models and even LLMs. By limiting the

[6] http://scholar.google.com
[7] https://ieeexplore.ieee.org/Xplore/home.jsp
[8] http://dl.acm.org
[9] http://www.sciencedirect.com
[10] https://aclanthology.org
[11] https://dblp.uni-trier.de
[12] http://www.semantic-web-journal.net

scope to this period, we prioritize contemporary research and avoid potentially outdated methodologies.

Our inclusion criteria were as follows:

- Peer-reviewed publications in English that focus on MRE.
- Approaches and datasets published between 2018 and 2024.
- For datasets, we also briefly discussed popular datasets available before 2018.

We excluded publications if they fulfilled at least one criterion of the following:

- Assessment methodologies published as a poster or abstract.
- Approaches and datasets that target monolingual approaches.
- Publications without a methodology or framework for MRE.

E. SEARCH METHODOLOGY STEPS

We structured our systematic literature search for MRE approaches based on a seven-step procedure:

1) Apply keywords to the search engine using the time-frame 2018–2024.
2) Scan article titles, and keywords based on our criteria.
3) Remove duplicates.
4) Review papers abstract based on our criteria.
5) Articles that met our criteria were thoroughly analyzed in relation to the research questions.
6) Retrieve new papers from the list of references cited by the papers of step 5.
7) Scan the references from the survey papers that passed steps 5 and 6 and retrieve additional papers that fulfill the criteria.

TABLE 3. Number of retrieved papers for each phase of the search methodology.

Search Engines	Steps						
	1	2	3	4	5	6	7
Google Scholar	246	94	82	50	16	16	19
IEEE Xplore	40	9	9	7	1	2	2
ACM DL	33	20	17	7	4	4	5
Science Direct	39	18	17	7	2	2	2
ACL	87	3	3	2	2	2	3
DBLP	88	68	62	28	8	8	8
SWJ	0	0	0	0	0	0	0
Total	533	212	190	101	33	34	39

Additionally, we incorporated any new surveys discovered during our search steps. Table 3 presents the number of papers we retrieved at each phase of the search methodology. Due to the large number of results from Google Scholar, ACM Digital Library, and Science Direct, we obtain the top-1000 results as the max number of returned papers.

IV. LITERATURE ANALYSIS

This section presents and examines the research findings, focusing on the research questions through a comprehensive interpretation and analysis of the literature from six key perspectives: domains covered, MRE approaches, explored languages, reproducibility, datasets, and evaluation metrics. These aspects cover methodological and practical dimensions, ensuring a holistic review of the field. By focusing on these areas, we address the core elements that impact MRE performance and generalizability across languages, making them essential for understanding current advancements and challenges, while avoiding less relevant or peripheral factors.

Each perspective is further broken down into specific categories within the existing literature. We summarize current research trends, identify challenges and difficulties identified by scholars, and offer insights for future work in Section V. Rather than attempting to cover all aspects of each paper in a single section, we distribute the discussion across relevant categories. For instance, in Section IV-A, we first discuss the problems each paper addresses within specific domains. Subsequent sections will then explore the approaches used, and other relevant aspects, ensuring a thorough and systematic analysis. Figure 2 shows our categorized overview of 18 MRE papers. Tables 4 and 6 present 18 approaches and 21 datasets, respectively.

A. MULTILINGUAL RELATION EXTRACTION DOMAINS/TYPES

We categorize the available approaches based on the domains or types they cover. Creating completely distinct categories is challenging—since some approaches could fit into more than one category—we have made our best effort to indicate when a paper fits multiple categories, placing it in the one most relevant to its primary contribution. We begin by dividing all approaches into two major domains: (1) open RE and (2) closed RE. Subsequently, we further subdivide the available literature into more specific subcategories.

1) OPEN RELATION EXTRACTION

Open RE, as discussed in Section II, is the process of extracting all possible relations from a text without a predefined set of fixed relation. The open RE concept was initially proposed by Etzioni et al. [53]. Most of the approaches in the literature are monolingual, with only a few addressing multilingual open RE prior to our selected timeframe. For instance, Zhang et al. [70], [71] and Faruqui and Kumar [67] explored multilingual open RE, primarily relying on language-dependent features and translation-based approaches. Faruqui and Kumar [67] made the first attempt to adopt transfer learning for MRE. However, their approach partially leverages the semantic information embedded in text.

Open RE presents a greater challenge than closed RE, as it requires identifying and extracting relationships without

TABLE 4. Approaches with their publication year, supported languages, number of relations, domain, dataset names, key performance indicators (KPIs), and implementation.

Approach	Year	Langs	#Rel	Domain	Dataset	KPI	Implementation
FA4RC [50]	2018	en, zh	6	Adversarial RE	ACE05	F1-score	https://github.com/zoubowei/feature_adaptation4RC
AMNRE [12]	2018	en, zh	176	Adversarial RE	Lin et al. [11]	AUC, Precision, Recall	https://github.com/thunlp/AMNRE
Structure Transfer [65]	2019	ar, en, zh	18	Joint Event and RE	ACE05	F1-score	-
NCL RE [68]	2019	ar, de, en, es, it, ja, pt, zh	59	MRE	In house & ACE05	F1-score	-
LOREM [60]	2020	en, es, fr, hi, it, nl, ru	-	Open RE	WMORC	Precision, Recall, F1-score	https://github.com/tomharting/LOREM
Multi^{2}OIE [61]	2020	en, es, pt	-	Open IE	Re-OIE2016	AUC, Precision, Recall, F1-score	https://github.com/youngbin-ro/Multi2OIE
GATE [30]	2021	ar, en, zh	18	Joint Event and RE	ACE05	F1-score	https://github.com/wasiahmad/GATE
LOME [63]	2021	en, zh	4	Temporal RE	Time Bank	F1-score	https://nlp.jhu.edu/demos/lome
PUCRJ-PUCPR-UFMG [62]	2021	en, es	13	Bio-Medical	eHealth-KD	Precision, Recall, F1-score	https://github.com/eHealth-KD-PUCs-UFMG/pucrj-pucpr-ufmg
CLARE [38]	2021	ar, en, zh	6	Adversarial RE	ACE05	F1-score	-
HERBERTa [16]	2021	ar, de, en, es, fa, fr, it, ko, nl, pl, pt, ru, sv, uk	16	Joint NER & RE	SMiLER	F1-score	https://github.com/samsungnlp/smiler
PARE [36]	2022	de, en, es, fr	37	Distant Supervision	Dis-ReX	AUC, F1-score	https://github.com/dair-iitd/DSRE
Temp Prob [64]	2022	en, es, fr, it	13	Temporal RE	Time Bank	F1-score	https://github.com/irenedini/tlink_probing
MRC Prompt [33]	2022	ar, de, en, es, fa, fr, it, ko, nl, pl, pt, ru, sv, uk	36	LLM Based RE	SMiLER	F1-score	https://github.com/DFKI-NLP/mefi-prompt
TransRel [69]	2023	bn, en, hi, te	51	Low resource	IndoRE	F1-score	https://github.com/NLPatCNERG/IndoRE
mERE [17]	2023	ar, de, en, es, fa, fr, it, ko, nl, pl, pt, ru, sv, uk	16	Joint NER & RE	SMiLER	F1-score	-
Prompt-XRE [39]	2023	ar, en, zh	18	Typed Open RE	ACE05 WMT17-EnZh XRE	F1-score	https://github.com/HSU-CHIA-MING/Prompt-XRE
SSDN [66]	2024	ar, en, zh	18	Joint EARL & RE	ACE05	F1-score	-

predefined labels. Unlike closed RE, which relies on a fixed set of relation types, open RE must dynamically infer relationships from various language contexts. This demands more sophisticated models capable of handling diverse linguistic structures and semantics [72]. The practical significance of open RE lies in its applicability to open-domain tasks, where predefined relations are impractical or insufficient. A prominent example is large-scale knowledge base construction from unstructured web data Stanovsky et al. [73].

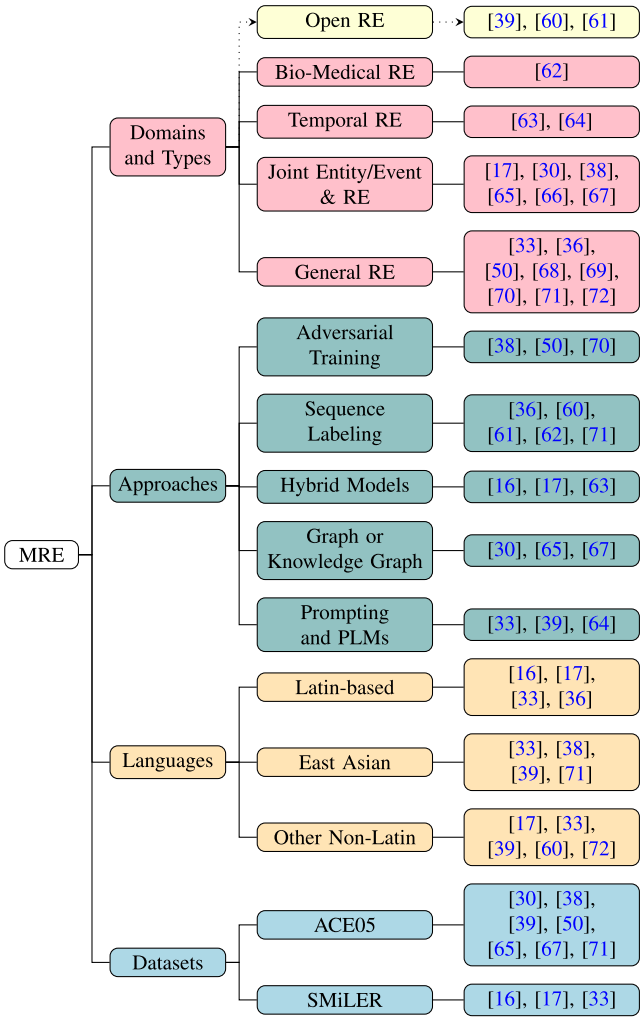


FIGURE 2. Our overview of 18 multilingual relation extraction papers categorized by domains and types, approaches, languages, and datasets.

In our literature survey, we identified three papers that address open RE in multilingual contexts. LOREM, introduced by Harting et al. [60], focuses on separately targeting language-independent and language-dependent features to eliminate the need for translation or external NLP tools such as POS taggers. Ro et al. [61] critique the reliance on LSTM models and monolingual embeddings in existing open RE models, proposing the use of multilingual BERT to outperform state-of-the-art approaches. The most recent work we found in the area of open RE is by Hsu et al. [39], who explored the use of large language models (LLMs) for RE. Although LLMs have shown superior performance in most NLP tasks, they had not been specifically focused on MRE until this study. The authors explored prompt tuning for MRE, which is categorized as a type-based open RE approach, does not limit itself to a fixed set of relations but rather provides sample relations for extraction. This approach can be seen as a bridge between open RE and closed RE, and given the increasing use of language models, it is likely to gain more popularity in the future.

2) CLOSED RELATION EXTRACTION

In contrast to open RE, closed RE involves the model predicting a predefined set of relations, typically limited to the relations on which the model was trained. Closed RE is important because it enables structured information extraction within a defined schema Shwartz and Dagan [74], which is particularly useful for domain-specific applications such as knowledge base population Ji et al. [75], question answering, and business intelligence. However, this approach comes with challenges. One major difficulty compared to open RE is its reliance on comprehensive and high-quality labeled data for training, which can be expensive and time-consuming to produce. Moreover, closed RE models often struggle with generalizing to unseen relations outside their predefined set, limiting their flexibility.

Closed RE is generally more common than open RE, for example, closed RE is used in financial document analysis to identify predefined relationships like `company-acquires-company` or `product-launchdate`. Similarly, it is employed in biomedical research for extracting relationships such as `gene-causes-disease`, aiding in knowledge discovery and hypothesis generation Wei et al. [76]. We further subdivide this category based on the domains and types discussed in the literature.

a: TEMPORAL RELATION EXTRACTION

Temporal RE identifies the sequence of events in text (e.g., `Before`, `After` and `Overlaps`) and is crucial for a variety of tasks such as structuring clinical data [77], text summarization [78], and question answering [79]. Temporal RE is a challenging task due to the complexity of identifying and linking temporally dependent events across diverse sentence structures and contexts. Natural language's inherent ambiguity in expressing time intervals and the domain-specific nature of temporal dependencies further complicate the task. These factors necessitate advanced models capable of handling various linguistic expressions and generalizing across domains Ning et al. [80]. In our literature review, we identified two key works discussing temporal RE. Xia et al. [63] presented a system that extracts an entity- and event-centric knowledge graph from textual documents. Rather than proposing a novel approach, they aimed to address a gap by presenting a multilingual system capable of efficiently extracting temporal relations between events and entities. Caselli et al. [64] investigated the use of LLMs for multilingual temporal RE. Their study offers a comprehensive analysis of the capabilities of multilingual LLMs, illustrating their impact on temporal RE.

b: BIO-MEDICAL RELATION EXTRACTION

Biomedical RE involves identifying and extracting relationships between biomedical entities (e.g., genes, proteins, diseases, drugs) from unstructured text sources such as scientific articles or clinical records [81]. Multilingual

biomedical RE is gaining importance as it enables the extraction of knowledge from a broader range of sources written in various languages, enriching biomedical databases and advancing global healthcare applications. Despite the dominance of English-based research, there is a noticeable scarcity of studies addressing multilingual biomedical RE, highlighting the need for further exploration in this area.

In our literature review, we identified only one related work that specifically addresses multilingual biomedical RE. In this study, Pavanelli et al. [62] developed a joint multilingual NER and RE system for biomedical domains. Their work addresses a critical gap in the field—the lack of multilingual biomedical information systems capable of handling diverse languages. This contribution is particularly valuable given the increasing need for extracting biomedical knowledge from non-English sources, which can significantly enhance the global accessibility and applicability of biomedical data. However, more research is needed to explore and develop robust multilingual systems that can further extend the capabilities of RE across a broader spectrum of languages and biomedical domain.

c: JOINT ENTITY/EVENT & RELATION EXTRACTION

This section addresses both multilingual entity and RE as well as multilingual event and RE, which do not fit into any specific domain. While these approaches are part of general MRE (Section IV-A2d), we categorize them separately due to their joint focus on both entity/event and RE.

A joint extraction model is designed to simultaneously extract multiple types of information from text, such as entities and their relationships, in a single unified framework. This contrasts with pipeline approaches that handle these tasks separately [17], [82]. Formally, let X be the input text, Y_1 and Y_2 be the entities and relations, respectively. The joint extraction model predicts both Y_1 and Y_2 simultaneously:

$$\mathrm{P}(Y_1, Y_2 \mid X) = \mathrm{P}(Y_1 \mid X) \cdot \mathrm{P}(Y_2 \mid X, Y_1) \tag{1}$$

Subburathinam et al. [65] address the challenge of multilingual event and RE by leveraging shared features across different languages. Previous work relied heavily on the distributional context of words within sentences, which did not effectively capture shared features across languages. To address this limitation, the authors use graph structures—such as constituency trees, dependency trees, and Part of Speech (POS) tagging—which are largely similar across languages. These structures enable the transfer of shared features from high-resource languages to low-resource languages, enhancing the extraction process. Later, Ahmad et al. [30] build upon this work by using an attention mechanism to overcome the issue of long-range dependencies and words that are not directly connected, thus improving the extraction of multilingual events and relations. Their approach addresses the limitations of previous models in handling distant relationships between words across languages.

In addition to work on joint multilingual event and RE, Yu et al. [38] tackle a common challenge in low-resource languages, where annotations are typically generated through translation-based methods. These methods often depend on machine translation models that fail when encountering unseen languages or words. To mitigate this, the authors use knowledge acquisition techniques tailored to multilingual entity and RE, thereby reducing reliance on translations. Seganti et al. [16] contribute to this area by presenting a multilingual dataset for joint entity and RE. They use distant supervision from Wikipedia and DBpedia to create this dataset and train a model called HERBERTa for multilingual IE. This work addresses the issue of the lack of large datasets for multilingual IE tasks. Wang et al. [17] introduce a novel approach to overcome the interference that can occur when transferring knowledge from high-resource languages to low-resource languages. In multilingual models, features from one language can sometimes interfere with those of another, leading to errors. To resolve this, the authors propose a language-independent switch that mitigates cross-linguistic interference and improves the performance of MRE. More recently, Wei et al. [66] addressed the challenges of MRE and Event Argument Role labelling (EARL) by incorporating semantic information alongside syntactic features. They argued that relying solely on syntactic information is insufficient in cross-lingual scenarios.

d: GENERAL RELATION EXTRACTION

We referred to this category as general RE because it encompasses the remaining literature we surveyed, which does not align with the previously discussed domains or types but focuses on closed MRE.

Lin et al. [11] developed a multilingual neural RE model, aiming to jointly represent texts from multiple languages to enhance RE performance. Building on this, Wang et al. [12] proposed a novel neural framework that explicitly encodes language consistency and diversity into different semantic spaces, thereby achieving more effective representations for MRE. While models trained on high-resource languages generally perform better, transferring latent features to low-resource languages remains challenging. To address this, Zou et al. [50] employed a generative adversarial network (GAN) to successfully transfer knowledge from models trained on high-resource languages to low-resource languages, thereby improving performance in low-resource settings. A well-known challenge in MRE is the lack of labeled data for low-resource languages, which makes it difficult to apply RE models across multiple languages. This problem is especially pronounced in languages with limited linguistic resources, such as annotated corpora, dependency parsers, or POS taggers. Ni and Florian [68] proposed a solution that allows for the use of an English-trained model in other languages with minimal resources. This method overcomes the barrier of linguistic resource scarcity, making MRE feasible even for low-resource languages. Rathore et al. [36] addressed a common issue in MRE approaches—many methods independently consider sentences during embedding, particularly in distant supervision-based systems. They introduced an

approach that takes into account all sentences available in a "bag" and generates contextualized embeddings using mBERT, allowing for better consideration of context in MRE tasks. With the rise of LLMs producing significant improvements in various NLP tasks, Chen et al. [33] explored whether prompting should be conducted in English or the target language, and whether to use soft prompt tokens for MRE. They also investigated how prompts perform under different learning scenarios—fully supervised, few-shot, and zero-shot learning. More recently, MRE research has focused on languages with extremely limited resources, such as Indian languages, which have seen almost no prior work in this area. Nag et al. [69] introduced the IndoRE dataset and applied a transfer learning model to achieve state-of-the-art performance for these underrepresented languages.

Domain-specific MRE, particularly in fields such as biomedicine, law, and finance, presents another set of challenges that differ from general-purpose text understanding. These domains often feature highly specialized vocabulary, long and syntactically complex sentences, and implicit relations that require deep contextual or world knowledge [83], [84]. For example, biomedical texts may include chemical or gene names that follow non-standard naming conventions, while legal documents may contain archaic or formal language not typically encountered in everyday NLP corpora. Additionally, domain-specific datasets are often costly to annotate due to the need for expert knowledge, resulting in small or imbalanced training sets. This data scarcity hinders model robustness and generalizability. Furthermore, the structure and discourse in such domains often diverge from conventional sentence-level relations, requiring models to handle document-level inference and cross-sentence relation extraction. These complexities call for tailored architectures, domain-adaptive pretraining, and specialized evaluation metrics to ensure effectiveness in real-world applications.

B. MULTILINGUAL RELATION EXTRACTION APPROACHES

We categorize the papers based on their underlying approaches and provide a concise description of each approach. Our focus here is not on the problems these studies addressed, as that was covered in the previous subsection. Instead, we highlight the methodologies and techniques employed. For each approach category, we provide a clear definition, representative methods, and a discussion of strengths and limitations to facilitate comparison. We divide the approaches into the following main subcategories and, where applicable, further subdivide them for a more detailed analysis.

1) ADVERSARIAL TRAINING

Adversarial training [85] is a technique used to improve the robustness of models. This is particularly beneficial for handling unseen or noisy data, where the input might deviate from the training examples. By training on adversarial examples—intentionally modified inputs designed to deceive the model, models become more robust and can accurately extract relations even from modified or slightly different text. Adversarial training has shown significant progress in machine learning and NLP tasks, leading researchers to explore its application in MRE. It aims to solve the following min-max optimization problem:

$$\min_{\theta} \max_{\delta \in \mathcal{S}} \mathcal{L}(f_\theta(x+\delta), y) \tag{2}$$

where $\mathcal{S}$ is the set of allowable perturbations, δ is the adversarial perturbation, f_θ is the model, and $\mathcal{L}$ is the loss function. In RE, this is crucial, as it helps models generalize better by making them resilient to minor text variations like synonym substitutions or rephrased sentences [86]. The primary objective of adversarial training in these studies is to achieve feature fusion, ensuring consistency across different languages.

Strengths: Adversarial training approaches excel in handling noisy or unseen data and ensuring language-consistent features across different languages. They are particularly effective for cross-lingual transfer where linguistic patterns vary significantly.

Limitations: These approaches require careful optimization and orthogonality constraints to avoid feature collapse. They also tend to be computationally intensive and may struggle with extremely low-resource languages where even adversarial examples are limited.

In our literature review, one of the early works using adversarial training is by Wang et al.'s model [12] that builds individual representations (vector representation) for each sentence to capture its unique linguistic features. It also constructs a consistent representation to encode shared features across languages. Adversarial training is employed here to capture language-consistent relation patterns from the consistent representations. To enhance the distinction between the individual and consistent representations, orthogonality constraints are introduced, ensuring that these two types of representations remain separate and complementary. Another notable study by Zou et al. [50] introduced an adversarial feature adaptation approach for cross-lingual RC. Their method uses a GAN to transfer feature representations from languages with rich annotations to those with limited labeled data. This approach demonstrated a marked improvement over existing techniques, particularly in under-resourced languages. Further advancements are seen in the work of Yu et al.'s CLARE [38] framework for cross-lingual RE. CLARE operates through a two-step process: Cross-Lingual Parallel Corpus Acquisition and Adversarial Adaptation and RE. The first module constructs a bilingual lexicon and translates the source language corpus into the target language while preserving the entity relationships. The second module then employs bilingual word embeddings and adversarial training to further improve cross-lingual RE, making the framework more robust for multilingual applications.

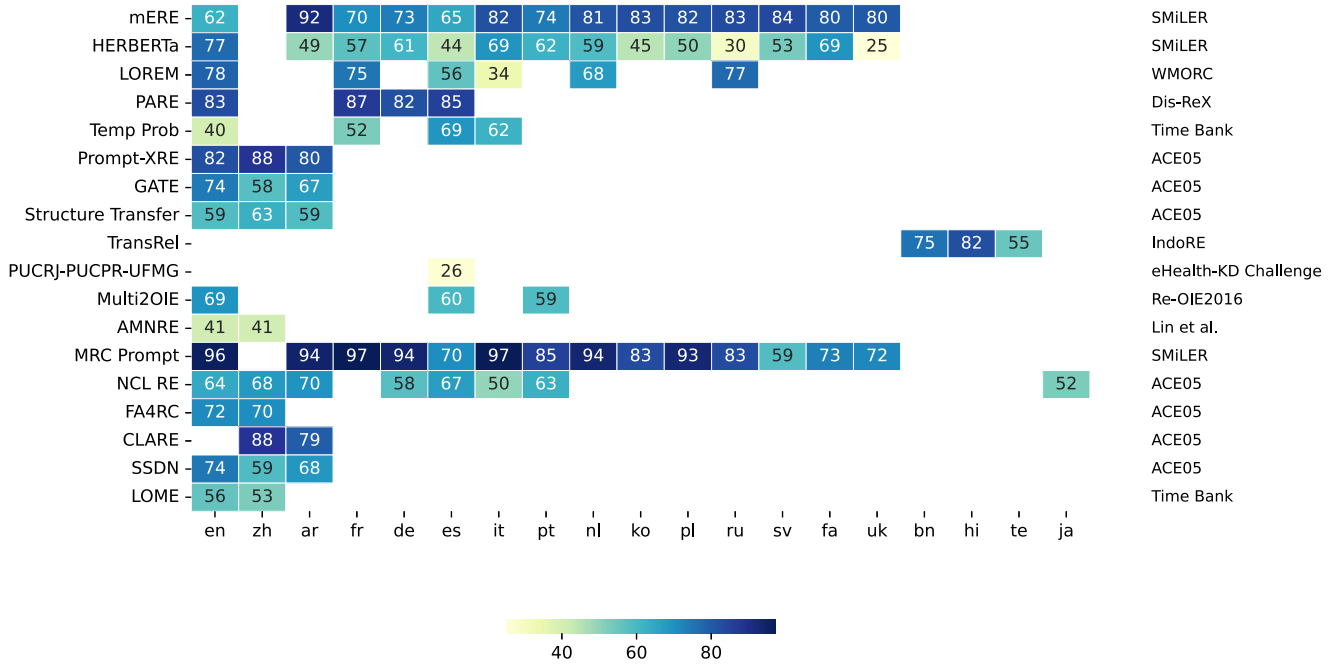


FIGURE 3. Performance of various approaches across languages. Empty cells indicate the unavailability of a particular language for that approach. The labels on the right represent the datasets used for evaluation.

2) SEQUENCE LABELING APPROACHES

Sequence labeling is the task of assigning a categorical label to each element in a sequence of observations. It is commonly used in NLP for tasks like POS tagging, NER, and chunking [87].

Formally, given a sequence of observations $X = \{x_1, x_2, \ldots, x_n\}$, the goal is to predict a sequence of labels $Y = \{y_1, y_2, \ldots, y_n\}$, where y_i is the label assigned to the observation x_i.

Sequence labeling techniques in RE primarily involve identifying and categorizing sequences within sentences to extract relational information [88], [89]. These methods often leverage pre-trained language models and multilingual embeddings to improve cross-lingual capabilities [90].

Strengths: Sequence labeling approaches are effective for fine-grained extraction tasks, particularly when leveraging multilingual embeddings. They can handle sentence-level extraction with high precision and are well-suited for identifying entity boundaries and relation spans simultaneously.

Limitations: These approaches often struggle with complex sentences containing multiple relations or long-distance dependencies. They may also face challenges with languages that have significantly different word order or morphological structures.

Ni and Florian [68] present a neural cross-lingual RE approach that focuses on improving the accuracy and generalization of extracting relations across multiple languages. It introduces a method leveraging multilingual word embeddings and transfer learning to address the scarcity of annotated data in non-English languages. They demonstrate the effectiveness of the approach on several datasets, showing significant improvements in performance across different languages. Multi^{2}OIE introduced by Ro et al. [61], is a multilingual OIE approach that utilizes multilingual BERT (mBERT) combined with multi-head attention blocks to extract relational tuples from sentences. This sequence-labeling system focuses on identifying all relations and extracting associated arguments. It leverages multi-head attention for relation token representation, enhancing the prediction of subject and object arguments. Furthermore, LOREM by Harting et al. [60] relies on monolingual open RE training data and pre-trained multilingual word embeddings. This sequence-labeling system integrates language-specific models with a language-consistent model trained on all available languages, assuming consistent relation patterns across languages. Pavanelli et al. [62] target the extraction of entities and relationships from clinical and biomedical texts evaluated in the eHealth-KD Challenge 2021. This approach utilizes mBERT to capture global dependencies within texts and to combine the entities' information. Passage-Attended RE (PARE) is another model introduced by Rathore et al. [36]. PARE processes all sentences mentioning an entity pair as a single passage, utilizing BERT (or mBERT for multilingual settings) to encode the entire passage. By employing an attention mechanism with the candidate relation as the query, PARE predicts relations more accurately.

3) HYBRID MODELS

Hybrid models combine various techniques, such as sequence classification and entity tagging, to perform RE in multilingual contexts.

Strengths: Hybrid models combine classification and tagging for improved performance in joint extraction tasks. They can leverage the advantages of multiple approaches and are often more flexible in handling diverse linguistic phenomena.

Limitations: These models face challenges with scalability and handling dense relation structures. They may also be more complex to implement and tune, requiring careful integration of different components.

The first hybrid model we discuss is HERBERTa presented by Seganti et al. [16]. This hybrid model first classifies an input sequence to identify relations, and then tags entities based on the classified relations. Evaluated on the SMiLER dataset across 14 languages, HERBERTa demonstrates effective joint extraction capabilities, but struggles with sentences containing multiple relations or entities. Xia et al. [63] introduce LOME, a large ontology multilingual extraction approach using multilingual encoders like XLM-R and multilingual training data. LOME performs coreference resolution, fine-grained entity typing, and temporal relation prediction. For MRE, it employs multilingual transfer learning and an SVM model. The mERE framework by Wang et al. [17] aims to improve multilingual entity and RE by addressing language interference through a two-stage training process. The Language-universal Aggregator captures shared features across languages, while the Language-specific Switcher refines these features for individual languages. For very low-resource languages, the authors Nag et al. [69] use ensemble learning to transfer knowledge from a high resource language to a low resource language.

4) GRAPH OR KNOWLEDGE GRAPH

Graph-based approaches in machine learning and NLP involve using graph structures to represent data and leveraging algorithms that operate on these graphs. These approaches are particularly useful for tasks involving relational data [91].

Strengths: Graph-based approaches are well-suited for representing relational data and addressing syntactic dependencies. They can effectively capture structural information that may be preserved across languages, making them valuable for cross-lingual transfer.

Limitations: These approaches can suffer from tokenization errors in languages with complex scripts and require robust graph construction techniques. They may also be computationally expensive for large-scale applications.

Ahmad et al. [30] propose a Graph Attention Transformer Encoder (GATE) that integrates syntactic information via self-attention mechanisms to capture relationships between non-adjacent words. Its reliance on syntactic dependencies enables robust language-agnostic representations, improving cross-lingual transferability. Subburathinam et al. [65] explore techniques for transferring relation and event extraction capabilities across languages without target language training data. The approach leverages Graph Convolutional Networks (GCNs) with symbolic and distributional features to construct a shared multilingual semantic space. Despite achieving comparable performance to supervised models, it encounters tokenization errors, particularly in Arabic and Chinese. The most recent advancement in this category by Wei et al. [66] utilizes a Semantic-Relation Graph Convolution Network to integrate semantic dependencies by combining these dependencies with syntactic information. They achieve enhanced performance in a cross-lingual setup, surpassing the results of Ahmad et al. [30] on the ACE05 dataset across three languages: Arabic, Chinese, and English.

5) PROMPTING AND PRE-TRAINED LANGUAGE MODELS

Prompting refers to the technique of guiding pre-trained language models by providing specific instructions or contextual information to elicit desired responses. It involves designing prompts that help the model generate text relevant to a particular task [92]. Prompting can be formalized as: $y = f_\theta(p \oplus x)$, where f_θ is the language model with parameters θ, p is the prompt, and x is the input text.

Strengths: Prompting and pre-trained language model approaches demonstrate strong performance in low-resource and cross-lingual scenarios by leveraging contextual knowledge. They can be particularly effective for zero-shot and few-shot learning settings.

Limitations: These approaches depend on well-designed prompts and resource-intensive training, which can limit scalability. Performance may vary significantly based on prompt design, and they typically require large models with substantial computational resources.

The approach by Chen et al. [33] uses prompting with PLMs for multilingual RC. Constructing prompts from relation triples with minimal translation for class labels. The results indicate optimal performance when prompting in the target language for supervised data and prompting in English instead of the target language for zero-shot scenarios. Caselli et al. [64] investigate temporal knowledge in the multilingual language model XLM-R compared to monolingual static embeddings, and used the contextualized knowledge of the PLM to know its ability for temporal relations. Evaluated on five temporally annotated corpora across four languages, it focuses on classifying temporal relations between event pairs. Prompt-XRE by Hsu et al. [39] addresses cross-lingual RE in low-resource languages using prompt-learning techniques. Prompt-XRE employs multilingual PLMs like mBART, utilizing hard, soft, and hybrid prompts for knowledge transfer across languages without target language labeling.

The discussed approaches in MRE offer unique strengths and limitations suited to different challenges. When selecting an approach, researchers should consider factors such as language resource availability, computational constraints, and the specific requirements of their application domain. For instance, adversarial training may be preferred for scenarios with significant linguistic divergence, while prompting approaches might be more suitable for extremely

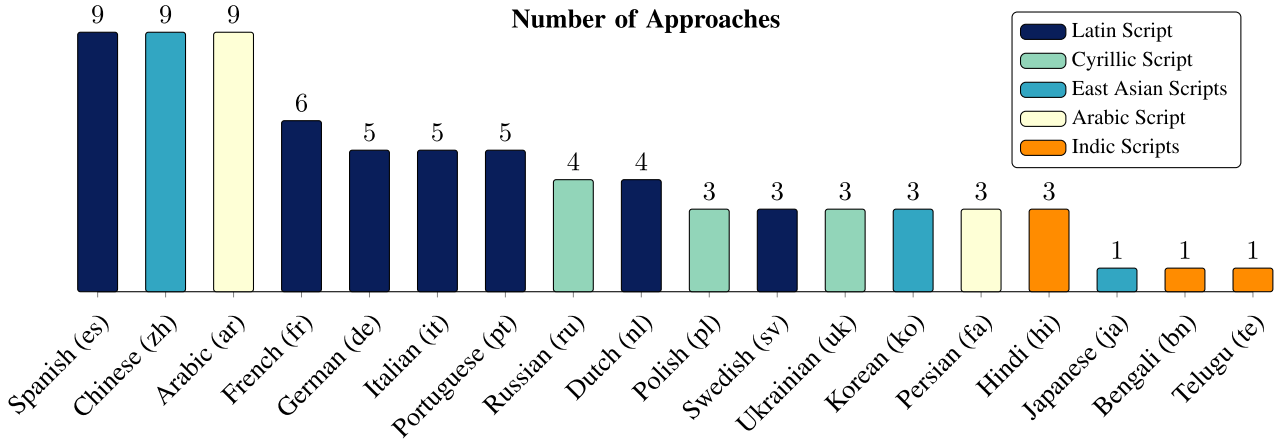


FIGURE 4. Frequency of languages across different approaches, categorized by language script. Bars represent the number of approaches supporting each language. Languages are grouped into five categories based on their writing scripts: **Latin script**, **Cyrillic script**, **East Asian scripts**, **Arabic script**, and **Indic scripts**. The y-axis indicates the number of approaches discussing each language, while the x-axis lists the languages covered.

low-resource languages where traditional supervised learning is infeasible.

C. LANGUAGES

We categorize languages based on their writing systems (scripts), which are highly relevant in MRE research due to their influence on tokenization, language model compatibility, and representation learning. Figure 4 illustrates the frequency of languages in MRE studies.

1) LATIN SCRIPT LANGUAGES

Languages such as English (en), French (fr), Spanish (es), Portuguese (pt), Italian (it), Dutch (nl), German (de), and Swedish (sv) use the Latin alphabet. These languages are the most widely studied in MRE, primarily due to extensive digital resources, rich annotation datasets (e.g., SMiLER, DIS-REx), and pre-trained language models trained on Latin-script corpora [93].

For instance, nine out of the 18 approaches report results for Spanish. Rathore et al.'s PARE [36] achieves the highest micro F1-score for Spanish using the DIS-REx dataset. In addition to Spanish, PARE also reports competitive results for French and German, where it attains the second-highest performance.

MRC-Prompt by Chen et al. [33] achieves the best micro F1-score on the SMiLER dataset for most Latin-script languages, including French (fr), German (de), Italian (it), Portuguese (pt), Dutch (nl), and Polish (pl). For Swedish (sv), MRC-Prompt's performance is comparable to that of mERE [17].

2) CYRILLIC SCRIPT LANGUAGES

Russian (ru), Ukrainian (uk), and Bulgarian (bg) use the Cyrillic alphabet. These languages are moderately represented in MRE studies, often included in multilingual benchmarks like SMiLER. For Russian and Ukrainian, MRC-Prompt performs competitively, while mERE [17] achieves the highest score for Ukrainian (uk), indicating variations in model generalization across Cyrillic-script languages.

The presence of morphological richness and different subword units in Cyrillic scripts makes generalization from Latin-script-pretrained models less straightforward.

3) EAST ASIAN LOGOGRAPHIC AND SYLLABIC SCRIPTS

Chinese (zh), Japanese (ja), and Korean (ko) use logographic or syllabic scripts, requiring character-level tokenization. Chinese, using logograms (Han characters), is commonly included in MRE datasets like ACE05 [94]. Both Prompt-XRE by Hsu et al. [39] and CLARE by Yu et al. [38] achieve similar performance for Chinese using the ACE05 dataset.

Japanese is underrepresented; Ni and Florian's NCL RE [68] is the only approach that evaluates Japanese (ja), using an in-house dataset that is not publicly available. Unfortunately, their implementation is also not accessible.

For Korean (ko), MRC-Prompt and mERE report similar results on the SMiLER dataset [17].

4) ARABIC SCRIPT LANGUAGES

Languages such as Arabic (ar), Persian (fa), and Urdu (ur) use the Arabic script. These languages pose challenges related to script segmentation, ligatures, and morphological complexity. Arabic is moderately well-studied in MRE due to its inclusion in ACE05 and SMiLER. MRC-Prompt by Chen et al. [33] achieves the highest score for Arabic on the SMiLER dataset, while Prompt-XRE by Hsu et al. [39] performs best on the ACE05 dataset.

For Persian (fa), the only dataset available is SMiLER, where mERE [17] achieves the highest score. These results show the importance of dataset-specific evaluations for Arabic-script languages.

5) INDIC SCRIPTS

Indic languages such as Hindi (hi), Bengali (bn), and Telugu (te) use Brahmic-derived scripts (e.g., Devanagari for Hindi, Eastern Nagari for Bengali, Telugu script for Telugu). These scripts are syllabic-abugidas and often pose tokenization and model embedding challenges. Despite large speaker populations, these languages are underrepresented in MRE.

LOREM by Harting et al. [60] uses an open RE approach and reports a 71.9 F1-score for Hindi (hi). For classification-based MRE, TransRel by Nag et al. [69] reports the highest scores for Hindi (hi), Bengali (bn), and Telugu (te). Notably, TransRel is the only approach that provides MRE datasets and models for both Telugu and Bengali, highlighting the severe resource limitations for many Indic languages.

6) OTHER SCRIPTS AND UNDERREPRESENTED REGIONS

African, Pacific Island, and indigenous American languages are almost entirely absent from current MRE research. Many of these languages use non-Latin scripts or have primarily oral traditions. This underrepresentation presents a major limitation for multilingual generalizability, highlighting the need for inclusive dataset development and script-aware modeling [17], [33].

Low-resource languages, such as many spoken in Africa (e.g., Yoruba, Amharic) and South Asia (e.g., Marathi, Sinhala) pose significant challenges for machine reading comprehension (MRC) and machine reading for MRE tasks. These languages often lack large-scale annotated datasets, high-quality corpora for pretraining, and even foundational NLP tools such as tokenizers, part-of-speech taggers, or dependency parsers [93], [95]. The scarcity of linguistic resources limits the training of language models that can understand nuanced grammatical structures, especially in morphologically rich and agglutinative languages. In addition, the diversity across dialects, non-standardized orthographies, and code-switching tendencies further complicate model generalization. Transfer learning from high-resource languages often performs suboptimally due to distributional mismatches and syntactic divergence. As a result, even multilingual pre-trained models like mBERT or XLM-R struggle with performance consistency across these languages, making low-resource MRE an open and critical research area.

D. REPRODUCIBILITY ANALYSIS

Reproducibility is a critical aspect of MRE research that directly impacts the field's progress. Our analysis reveals significant challenges in reproducing existing MRE methods, primarily due to the lack of standardized evaluation protocols, incomplete reporting of experimental details, and limited availability of source code and pre-trained models.

Table 5 presents a comprehensive assessment of reproducibility factors across the surveyed approaches. We evaluate each approach based on four key criteria: (1) availability of source code, (2) provision of pre-trained models, (3) documentation of data splits, and (4) completeness of implementation details.

Our analysis reveals several key findings:

1) Only 72% of approaches provide publicly available source code, creating significant barriers to reproduction.
2) Pre-trained models are available for only 22% of approaches, forcing researchers to retrain models from scratch, often with insufficient details.
3) Data splits are documented for 50% of approaches, making direct performance comparisons challenging.
4) Complete implementation details are provided for 61% of approaches, with the remainder offering only partial or minimal information.

TABLE 5. Reproducibility assessment of MRE approaches.

Approach	Reproducibility Aspects			
	Code	PLM	Data Splits	Imp. Details
FA4RC	✓	✗	✗	Partial
AMNRE	✓	✗	✗	Partial
Structure Transfer	✗	✗	✗	Partial
NCL RE	✗	✗	✗	Minimal
LOREM	✓	✗	✓	Complete
Multi2OIE	✓	✗	✓	Complete
GATE	✓	✗	✗	Partial
LOME	✓	✓	✗	Complete
PUCRJ-PUCPR-UFMG	✓	✗	✓	Complete
CLARE	✗	✗	✗	Partial
HERBERTa	✓	✓	✓	Complete
PARE	✓	✗	✓	Complete
Temp Prob	✓	✗	✓	Complete
MRC Prompt	✓	✓	✓	Complete
TransRel	✓	✓	✓	Complete
mERE	✗	✗	✗	Partial
Prompt-XRE	✓	✗	✗	Complete
SSDN	✗	✗	✗	Partial

These findings highlight a critical need for improved reproducibility practices in MRE research. We recommend that future work in this area adopt standardized reporting practices, including:

1) Publishing source code with clear documentation
2) Providing pre-trained models or detailed training procedures
3) Clearly documenting data splits and preprocessing steps
4) Establishing unified benchmarks for consistent evaluation

Addressing these reproducibility challenges would accelerate progress in MRE research by enabling more effective comparison of approaches and facilitating incremental improvements to existing methods.

E. MULTILINGUAL RELATION EXTRACTION DATASETS

The landscape of MRE has seen substantial advancements with the introduction of several comprehensive datasets, each addressing unique challenges. As MRE approaches, such as those previously discussed, are generally assessed and evaluated using such benchmark datasets, this survey includes a comparison of these datasets.

We begin by outlining the important characteristics of a MRE dataset. We highlighted key features of a high quality MRE dataset. These features are summarized in Table 6 for each dataset discussed in the following sections. We categorize each dataset according to the following dataset criteria: human-annotated, machine translation, automatic model-annotated, and other annotation methods. Datasets that are based on machine translation but include human annotations are listed in the human-annotated datasets category.

1) IMPORTANT FEATURES

A MRE dataset should include a diverse range of languages from different families and regions, ensuring balanced examples and diverse relations across all languages [96]. Clear guidelines for the annotation process have to be provided and to be followed during the annotation. The annotations must be accurate, consistent, and verified by native speakers [97]. The dataset should offer rich context, including entire sentences or paragraphs, and document-level context if possible. Cross-lingual alignments and parallel corpora are crucial for facilitating transfer learning [98]. Texts from various domains and genres, both formal and informal, should be included for diversity [99], [100]. Additional linguistic features, metadata, and ethical considerations such as privacy, consent, and bias mitigation are essential [101]. The dataset should be publicly available with thorough documentation and standardized evaluation metrics [102]. Baseline models and results should be provided to facilitate benchmarking and comparison [103].

2) HUMAN-ANNOTATED DATASETS

MixRED [18] is a recent dataset[13] that introduces a novel mix-lingual RE dataset blending English and Chinese documents to address the challenge of RE in code-switching contexts. With diverse mixing strategies and comprehensive annotations, MixRED mitigates relation bias and allows for the exploration of multilingual patterns. Supervised models outperform LLMs on the dataset, highlighting the complexity of mix-lingual tasks, but pretrained mix-lingual patterns improve performance. While MixRED sets a strong foundation for MRE, challenges remain in effectively capturing mix-lingual dependencies.

The Multi-CrossRE [104] dataset[14] is a machine translated version of CrossRE [105] that includes 26 languages in addition to English, and covering six text domains for sentence-level RE. This dataset includes a portion of 200 sentences in seven languages checked by naive speakers including Czech, Danish, Dutch Finnish, German, Italian, and Japanese.

RED^{FM} and SRED^{FM} datasets[15] by Huguet Cabot et al. [9] introduce comprehensive resources for MRE, spanning 7 and 18 languages respectively. They include high-quality annotations achieved through a Transformer-based NER classifier and human reviews, improving multilingual representation and annotation quality. These datasets, while broadening the linguistic coverage seen in earlier datasets, highlight potential issues with balance across relations and languages.

Addressing domain-specific needs, the BIZREL [44] dataset[16] focuses on business RE across Chinese, English, French, and Spanish. Unlike SMiLER, which is primarily sourced from Wikipedia, BIZREL draws from diverse sources like online news and industry reports, offering contextually rich sentences. Despite a higher frequency of French and English instances, BIZREL's rigorous manual annotations and cross-lingual alignment mark a significant improvement in multilingual business RE.

The eHealth-KD Challenge [106] dataset[17] comprises English and Spanish texts and covers healthcare and news domains for the recognition of entities and their relations.

The Re-OIE2016 [107] dataset[18] was released in English for the OIE task and is the relabeled test dataset[19] of OIE2016 [73]. The Re-OIE2016 dataset was automatically translated and re-annotated by Ro et al. [61] for Portuguese and Spanish.

The WMORC [67] dataset[20] consists of two parts gathered from Wikipedia for the Open RE task. The first part contains manually annotated data for three languages: French, Hindi, and Russian. The second part contains automatically tagged data for 60 languages.

The ACE05 [94] benchmark is one of the first and most widely used RE dataset[21] that consists of human annotations of relations for three languages: Arabic, Chinese, and English but requires a paid license for its use. The text sources of the datasets are news, newsgroups, conversations, and weblogs. The datasets include 7 entity types and 18 relation subtypes.

The English TimeBank [108] dataset[22] consists of 183 English news articles with over 27,000 event and temporal annotations with 13 finegrained temporal values. The TimeBankDense [109] dataset approximates a complete graph of all possible temporal relations over events and temporal expressions from the training portion of the English TimeBank. This dataset contains only 36 documents and 5 temporal relations. The Spanish TimeBank [110] dataset[23] consists of 210 manually annotated documents with a simplified set of 5 temporal relations. The Italian TimeBank [111] dataset consists of 254 documents with 13 temporal values. The French TimeBank [112] dataset consists of 107 documents with 13 relations.

[13] https://github.com/acddca/MixRED
[14] https://github.com/mainlp/CrossRE
[15] https://github.com/babelscape/rebel
[16] https://github.com/Geotrend-research/business-relation-dataset
[17] https://github.com/ehealthkd/corpora
[18] https://github.com/zhanjunlang/Span_OIE
[19] https://github.com/gabrielStanovsky/supervised-oie
[20] https://www.kaggle.com/datasets/shankkumar/multilingualopenrelations15
[21] https://catalog.ldc.upenn.edu/LDC2006T06
[22] https://catalog.ldc.upenn.edu/LDC2006T08
[23] https://catalog.ldc.upenn.edu/LDC2012T12

The Multi-SimLex [46] dataset[24] and the MultiLexBATS [47] dataset[25] are for lexical semantic RE. Multi-SimLex is a lexical resource covering diverse monolingual and 66 cross-lingual datasets. The monolingual datasets provides human judgments for Arabic (ar), English (en), Estonian (es), Finnish (fi), French (fr), Hebrew (he), Kiswahili (sw), Mandarin Chinese (cmn), Polish (pl), Russian (ru), Spanish (sp), Welsh (cym), and Yue Chinese (yue). Each monolingual language dataset is annotated for the lexical relation of semantic similarity and contains 1,888 semantically aligned concept pairs. The MultiLexBATS dataset of lexical semantic relations comprises of translations to 15 languages (manually curated) of the English BATS [113] dataset and covers four groups of relations: inflexion morphology, derivational morphology, lexicographic semantics, encyclopedic semantics [114]. Each of these groups has relations, e.g., `hypernyms-animals`, `meronyms-substance`, and `synonyms-intensity`.

3) MACHINE TRANSLATION DATASETS

The MultiTACRED [10] dataset[26] extends the TACRED RE dataset into 12 typologically diverse languages, employing machine translation and automatic annotation projection. This dataset enables comprehensive evaluations across monolingual, cross-lingual, and multilingual models, maintaining high annotation quality despite facing translation errors and language-specific annotation issues. MultiTACRED significantly advances the field by providing a rich, diverse dataset supporting extensive linguistic and cross-lingual research.

Expanding the linguistic diversity further, the SMiLER [16] dataset[27] is an open-domain corpus of annotated sentences, created for the Joint Entity and RE task that incorporates six languages, including Korean and Portuguese, with meticulous annotation processes that combine automated methods and human validation. This dual approach ensures high annotation quality, providing a robust foundation for evaluating multilingual models. However, challenges in handling no_relation sentences and maintaining consistency in automated annotations persist, an issue also noted in the previous datasets.

Datasets that are based on machine translation but include human annotations are listed in the previous section.

4) AUTOMATIC MODEL-ANNOTATED DATASETS

The WMT17-EnZh XRE [39] dataset[28] is a cross-lingual RE dataset that contains 0.9M English-Chinese entity mention pairs automatically extracted from the WMT17 En-Zh parallel corpus [115]. Addressing the shortcomings of previous datasets, the IndoRE [69] dataset[29] provides a balanced resource focusing on Indian languages that encompasses Bengali, Hindi, Telugu, and English, addressing the morphological and syntactic diversity unique to these languages. With 21,000 entity- and relation-tagged sentences, IndoRE provides a robust testbed for low-resource language research, adhering to ethical guidelines and significantly advancing RE capabilities in Indian languages.

The DiS-ReX [116] dataset[30] provides a balanced and cross-lingually aligned resource spanning English, French, German, and Spanish. With over 1.8 million sentences from Wikipedia, DiS-ReX ensures diverse and contextually rich data, similar to the comprehensive coverage seen in X-WikiRE and EGD. However, DiS-ReX excels by addressing class imbalance and offering a realistic benchmark for distant supervision RE tasks.

The mSubEvent [117] dataset[31] is only accessible with a password. The dataset covers five languages, including less common ones like Danish and Urdu, with high-quality annotations leveraging Wikipedia articles. This dataset's approach of segmenting articles into manageable chunks for detailed annotation is similar to the thorough methods employed in SMiLER. However, it highlights the challenges in non-English language annotations, emphasizing the need for further research in this area.

The RELEX [118] dataset[32] introduces RELX and RELX-Distant, targeting cross-lingual RC with high-quality parallel sentences in multiple languages, including Turkish. While RELX ensures human-translated annotations, RELX-Distant employs distant supervision, akin to the methods seen in EGD, but potentially introduces noise. Both datasets contribute to robust cross-lingual NLP models, particularly for low-resource languages.

5) OTHER ANNOTATIONS

The GDS [119] dataset[33] Guided Distant Supervision for creating the largest German biographical RE dataset with over 80,000 instances and nine relation types. GDS improves label accuracy using resources like Pantheon and Wikidata, though challenges remain with less precise German entity recognition models and complex relation annotation. Cross-lingual learning between English and German models shows strong potential for low-resource languages, despite difficulties in classifying some relations, overall, the study offers valuable datasets and models for advancing MRE.

The EGD [120] dataset leverages event-guided denoising techniques to filter out low-quality examples from date-marked news articles, resulting in high-quality relation statements in English and Spanish. This approach significantly reduces training costs while maintaining competitive performance, offering a more resource-efficient alternative compared to X-WikiRE. However, its reliance on a large news corpus limits applicability in low-resource languages.

[24] https://multisimlex.com/#download
[25] https://github.com/nexuslinguarum/MultiLexBATS
[26] https://catalog.ldc.upenn.edu/LDC2018T24
[27] https://github.com/samsungnlp/smiler
[28] https://github.com/HSU-CHIA-MING/Prompt-XRE
[29] https://github.com/NLPatCNERG/IndoRE
[30] https://github.com/dair-iitd/DiS-ReX
[31] http://nlp.uoregon.edu/private/mSubEvent-v0.1
[32] https://github.com/boun-tabi/RELX
[33] https://huggingface.co/datasets/plumaj/biographical

Moreover, the authors have neither shared their source code nor data in their given repository.

The X-WikiRE [121] dataset stands out with its multilingual coverage, including languages such as English, German, French, Spanish, and Italian. Framed as a reading comprehension task, X-WikiRE enhances zero-shot learning capabilities and facilitates cross-lingual transfer with minimal target language fine-tuning, setting a new benchmark for MRE.

F. EVALUATION BENCHMARKS AND METRICS

In this section, we discuss the selected literature based on the most widely used benchmark datasets for evaluation. We also briefly review the evaluation metrics commonly applied, and the performance comparisons based on these metrics. The studied literature are grouped into three major categories according to their frequency of use: those using the ACE05 dataset by Walker and Consortium [94], the SMiLER dataset by Seganti et al. [16], and others. The ''Others'' category encompasses approaches that use datasets that are rarely used or only have one related study available. The details of properties and features of all the datasets including ACE05 and SMiLER are covered in the next section also their properties are given in 6. Figure 5 illustrates the performance of various approaches on their respective datasets.

1) SYSTEMS BASED ON THE ACE05 DATASET

Figure 5b shows the performance of different approaches across various languages using the ACE05 dataset introduced by Walker and Consortium [94]. Due to its extensive history of nearly 20 years, ACE05 is the most widely used dataset for MRE, with seven out of the 18 approaches in our study employing it. The Prompt-XRE approach by Hsu et al. [39] demonstrates superior performance across nearly all three languages (Arabic, Chinese, and English). One contributing factor to this success is the recent advancements in LLMs, which have significantly improved overall performance.

2) SYSTEMS BASED ON THE SMILER DATASET

Figure 5 depicts the performance of different approaches on the SMiLER dataset. Although SMiLER is a relatively new MRE dataset, its comprehensive coverage of multiple languages has made it a popular choice among researchers. Three out of the 18 studies evaluated their systems using this dataset. For Latin-based languages, the MRC-Prompt approach by Chen et al. [33] consistently outperforms others. However, for languages that deviate from Slavic script, such as Russian and Ukrainian, the mERE approach achieves better results.

3) METRICS

Different metrics are employed to evaluate various aspects of MRE system performance. The micro F1-score measures the overall system performance, prioritizing frequent relation types, making it suitable for imbalanced datasets. The macro F1-score, on the other hand, treats all relation types equally, assessing the model's ability to generalize to rarer relations. While F1-score balances precision and recall, both micro and macro F1-score are essential for a comprehensive evaluation, considering both frequent and rare classes. Additionally, the AUC metric complements F1-score by assessing the model's ability to discriminate between positive and negative instances across different thresholds, especially valuable for imbalanced scenarios. By combining these metrics, we can obtain a holistic evaluation of MRE systems.

From our review of the literature, we found that F1-score is the most commonly used evaluation metric, with two variants: micro F1-score and macro F1-score. In addition to F1-scores, some studies, such as LOREM [60], PUCRJ-PUCPR-UFMG [62], and Multi^{2}OIE [61], also report Precision and Recall values. Furthermore, PARE [36], Multi^{2}OIE [61], and AMNRE [12] provide the Area Under the Curve (AUC) metric. Notably, AMNRE does not evaluate its approach using F1-scores. In our analysis, we focus primarily on micro F1-scores, wherever applicable. In cases where F1-scores are not reported, such as with AMNRE, we calculate F1-scores using the reported Precision and Recall values. The definitions and mathematical formulations of these metrics are presented in Section II-L.

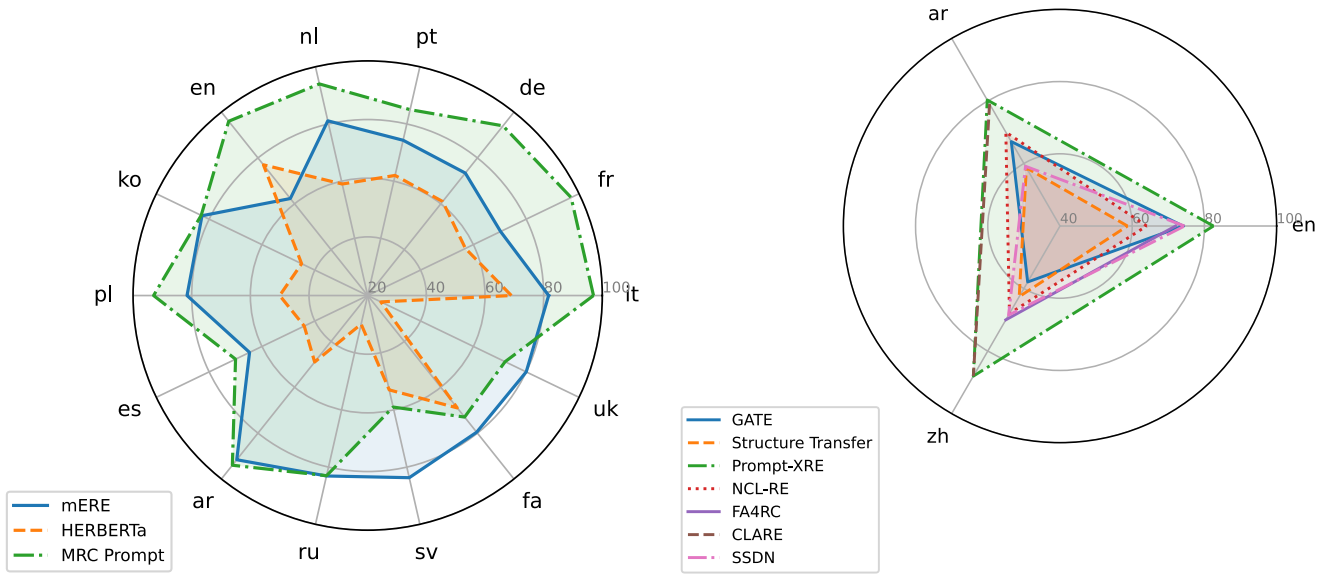
G. PERFORMANCE AND CHALLENGES

In this section, we review key MRE approaches in chronological order, focusing on challenges and performance-related insights. We highlight specific issues such as language diversity, dataset limitations, and evaluation gaps that persist across methods. Figure 3 presents the performance of each approach across different datasets in the form of a heat map.

We begin with AMNRE by Wang et al. [12], which achieved state-of-the-art results on 176 relations at the time of publication. While the code was made available, the evaluation omitted F1-score metrics and was restricted to high-resource languages such as Chinese and English. Similarly, the adversarial method proposed by Zou et al. [50] in the same year included F1-score evaluations and public code but also focused exclusively on high-resource language settings. These trends reflect a broader tendency among early approaches to prioritize strong performance on well-resourced languages while overlooking multilingual applicability.

In 2019, Ni and Florian [68] identified difficulties when handling syntactic structures unfamiliar to English-based models, such as SOV (Subject-Object-Verb) or VSO (Verb-Subject-Object) orders. Their model also struggled in the absence of bilingual embeddings or dictionaries. A related effort by Subburathinam et al. [65] failed to report competitive results or compare against strong baselines, making its effectiveness difficult to assess. These findings suggest that structural and lexical divergence across languages poses persistent challenges for cross-lingual MRE systems.

The multilingual open RE system LOREM by Harting et al. [60] demonstrated good performance for high-resource



(a) SMiLER dataset

(b) ACE05 dataset

FIGURE 5. Comparison of approach performance across different languages. The SMiLER chart compares three approaches across 14 languages. The ACE05 chart focuses on seven approaches and three languages, with 20-unit intervals on the radial axis for clarity.

languages but faltered in low-resource contexts, such as Hindi (with an F1-score of 0.054). Moreover, its ability to handle complex sentences with more than two entities was limited. In contrast, Multi^{2}OIE [61] performed robustly across languages—even without specific training on low-resource ones—demonstrating better generalization.

Approaches in 2021 continued to face similar limitations. CLARE [38] offered results only for Arabic and Chinese, with no reproducible implementation. PUCRJ-PUCPR-UFMG [62], focused on Spanish within the healthcare domain, achieved modest scores despite a 4th-place ranking in a shared task. These works underscore the difficulty of developing generalizable solutions when language and domain coverage remain narrow.

LOME [63] lacked publicly available code, and although it employed the TimeBank dataset, this resource was not multilingual by design. GATE [30] showed promise but was limited to sentence-level RE and encountered issues with entity ordering. HERBERTa [16] observed language-specific error behaviors and reported underwhelming performance on English. While data from their study was shared, model weights were not included. These studies highlight how dataset limitations and language-specific characteristics often constrain model performance and interpretability.

Prompting-based approaches began to gain traction in 2022, as demonstrated by Chen et al. [33], who achieved over 0.95 micro-F1-score in fully supervised setups and found that zero-shot prompting in English transferred best across languages. This signals a shift toward more effective multilingual generalization, albeit with lingering dependencies on dominant languages.

PARE [36] achieved moderate results (e.g., 4% macro F1-score, 3.2% micro) on the DISRex dataset but primarily targeted monolingual settings and did not use competitive baselines. The underlying dataset was a combination of separate monolingual corpora rather than a unified multilingual benchmark, limiting the scope of evaluation.

Recent works from 2023 continue this trajectory. Prompt-XRE [39] reported a 5% improvement in F1-score but did not include widely-used benchmarks like SMiLER in its evaluation. Meanwhile, mERE [17] achieved top F1-score results for 9 out of 14 languages on SMiLER but lacked public code and did not disentangle entity and RE in evaluation. TransRel [69] focused on extremely low-resource languages with an openly available dataset, though model checkpoints were not provided, making replication and comparison challenging.

Overall, while recent advances—particularly those using prompting and pre-trained models—demonstrate improved multilingual performance, issues such as dataset fragmentation, incomplete evaluation, and limited low-resource language coverage continue to restrict progress in MRE.

V. OPEN DIRECTIONS

Despite significant advancements, the methodologies currently employed in MRE face numerous challenges that necessitate targeted improvements. This section outlines key suggestions and improvements for existing methodologies, aiming to foster more robust and effective MRE systems.

TABLE 6. Multilingual relation extraction datasets with the release year (Year), the supported languages (Langs), whether the data set is balanced (Bal), how it is annotated (Annotation), the dataset source (Src), whether different sources exist (Var), whether a source or licences is given (E), the licence source (Access), proper evaluation of the dataset (Eval), and the number of relations (#Rel).

Dataset	Year	Langs	Bal	Annotation	Src	Var	E	Access	Eval	#Rel
MultiLexBATS [47]	2024	15 languages	✗	Human based	BATS [113] dataset	✗	✓	Open	✓	-
GDS [119]	2024	en,de	✗	Distant Supervision	Wikipedia/Wikidata Pantheon	✗	✓	Open	✓	9
MixRED [18]	2024	en, zh	✗	Human based	VOA news	✗	✓	Open	✓	21
RED^{FM} [9]	2023	ar, de, en, es, fr, it, zh	✗	Amazon Mechanical Turk	Wikipedia/Wikidata	✗	✓	Open	✓	400
Multi-CrossRE [104]	2023	27 languages	✗	Machine Translation Human based	artificial intelligence, literature, music, news, politics, natural science	✓	✓	Open	✓	17
MultiTACRED [10]	2023	ar, de, es, fi, fr, hi, hu, ja, pl, ru, tr, zh	✗	Machine Translation	TAC, KBP	✓	✓	Licensed	✓	41
WMT17-EnZh XRE [39]	2023	en, zh	✗	Automatic Model Based	WMT 2017 parallel corpus	✗	✗	Open	✗	-
IndoRE [69]	2023	bn, en, hi, te	✓	Automatic Model Based	Wikipedia/Wikidata	✗	✓	Open	✓	51
Multi-SimLex [46]	2020	ar, eng, es, fi fr, he, sw, cmn pl, ru, sp, cym, yue	✗	Human based	SimLex-999, SEMEVAL-500 CARD-660, SimVerb-3500,USF	✗	✓	Open	✓	-
DiS-ReX [116]	2022	de, en, es, fr	✓	Automatic Model based	Wikipedia/Dbpedia	✗	✗	Open	✓	37
mSubEvent [117]	2022	da, en, es, tr, ur	✗	Automatic Model based	Wikipedia	✗	✓	NA	✗	-
BIZREL [44]	2022	en, es, fr, zh	✗	Human based	NEWS, companies websites/reports	✓	✓	Open	✓	5
SMiLER [16]	2021	de, es, fr, it, ko, pt	✗	Translation based	Wikipedia	✗	✗	Open	✓	16
eHealth-KD [106]	2021	en, es	✗	Human based	electronic health documents, news	✗	✓	Open	✓	13
RELX [118]	2020	de, en, es, fr, tr	✓	Automatic Model based	KBP-37	✗	✗	Open	✓	37
Re-OIE2016 [107]	2020	en, es, pt	✗	Machine Translation Human based	WSJ and encyclopedia	✗	✓	Open	✓	-
EGD* [120]	2020	en, es	✗	Event-guided pairing	Reuters NEWS	✗	✗	Empty	✗	-
X-WikiRE [121]	2019	de, en, es, fr, it	✗	Reading comprehension	Wikipedia/Wikidata	✗	✗	NA	✓	-
WMORC [67]	2015	63 languages	✗	Human and automatically tagged	Wikipedia	✗	✓	Open	✓	-
ACE05 [94]	2005	ar, cmn, en	✗	Human based	broadcast conversation, broadcast news, newsgroups, telephone conversations, weblogs	✓	✓	Licensed	✓	18
TimeBank [108]	2003	en	✗	Human based	news articles	✗	✓	Licensed	✓	-

A. GENERAL OPEN DIRECTIONS

1) ENHANCEMENT OF DATA RESOURCES

One of the challenges in MRE is the lack of high-quality annotated datasets for low-resource languages. To address this, researchers have proposed several strategies.

a: CROWDSOURCING AND COMMUNITY ENGAGEMENT

Leveraging crowdsourcing platforms to gather annotations from native speakers can significantly enhance the volume and quality of training data. Engaging local linguistic communities not only aids in data collection but also ensures that the nuances of each language are accurately captured [20], [119].

b: SYNTHETIC DATA GENERATION

Employing techniques such as data augmentation and transfer learning can help generate synthetic training examples. By utilizing high-resource language datasets to create parallel corpora, researchers can improve the performance of MRE systems in low-resource contexts [122]. Specific techniques include back-translation with controlled noise, entity substitution preserving relation semantics, and template-based

generation with linguistic constraints tailored to target language morphology.

c: CROSS-LINGUAL TRANSFER LEARNING

Developing models that can effectively transfer knowledge from high-resource languages to low-resource languages is essential. This can be achieved by fine-tuning multilingual pre-trained models on specific language pairs, allowing for the adaptation of learned representations to new linguistic contexts [119], [123]. Recent advances in cross-lingual alignment techniques, such as contrastive learning with parallel data and feature-level orthogonality constraints, show particular promise for preserving relation semantics across language boundaries.

While synthetic data generation and cross-lingual transfer learning offer promising avenues for improving MRE, each approach is accompanied by significant technical challenges. Synthetic datasets, for example, often suffer from distributional differences compared to real-world data, leading to domain shift and reduced model generalization [124]. Automatically generated text may lack the subtle contextual cues or relational depth present in human-authored documents. To mitigate this, recent work has explored contrastive learning and adversarial training as mechanisms for aligning feature representations across synthetic and real domains [125]. Similarly, in MRE, transfer learning from high-resource to low-resource languages is frequently hindered by linguistic divergence, lack of parallel corpora, and noise introduced by machine translation [126]. Advances such as prompt-tuning [127] and adapter-based fine-tuning provide flexible, parameter-efficient alternatives that can be adapted to multilingual or domain-specific settings. Moving forward, combining these techniques with robust data selection and domain adaptation strategies will be critical for scaling MRE systems to real-world, diverse scenarios.

2) METHODOLOGICAL INNOVATIONS

The methodologies used in MRE can benefit by adopting several cutting-edge approaches as follows.

a: ADVANCED CONTEXTUAL AND CROSS-LINGUAL EMBEDDINGS

Leveraging transformer-based models like XLM-R and multilingual BERT has shown promise in capturing cross-lingual semantic relationships without explicit alignment [119]. Furthermore, integrating prompt-based learning frameworks can provide adaptable templates that enhance the generalization of RE across diverse languages and domains with minimal supervision.

b: HANDLING LONG DOCUMENT CONTEXTS

In scenarios involving long documents, employing models such as Longformer [128] or BigBird [129], which are designed to process extended text sequences, can preserve global context while accurately extracting complex relations. Hierarchical attention mechanisms can further refine this process by focusing on relevant document sections without losing the broader narrative. Document-level RE can be enhanced through multi-granularity modeling that captures entity interactions at sentence, paragraph, and document levels, with explicit modeling of coreference chains and discourse structures.

c: HIERARCHICAL AND GRAPH-BASED MODELS

Implementing hierarchical models that consider the structural relationships between entities can improve the extraction of complex relations [130]. Additionally, graph-based approaches that utilize KGs can facilitate the integration of external knowledge, providing a more comprehensive understanding of entity relationships [131], [132]. Graph neural networks with cross-lingual knowledge alignment mechanisms can bridge language-specific knowledge gaps by transferring relation patterns across languages with different structural properties. Specifically, heterogeneous graph attention networks that model both syntactic dependencies and semantic relationships show promise for capturing language-universal relation patterns.

3) EVALUATION AND BENCHMARKING

To ensure the effectiveness of MRE systems, robust evaluation frameworks must be established.

a: STANDARDIZED BENCHMARKS

Developing standardized benchmarks for MRE that include diverse languages and domains is crucial. This will allow for consistent evaluation and comparison of different methodologies, facilitating a more competitive research environment [133].

b: COMPREHENSIVE EVALUATION METRICS

Expanding the evaluation metrics beyond traditional accuracy and F1 scores to include metrics that assess the contextual relevance and cultural appropriateness of extracted relations can provide deeper insights into the performance of MRE systems. For example, incorporating user-centric evaluation metrics that reflect end-users' perspectives by conducting user studies where domain experts evaluate extracted relations based on their applicability and relevance to specific tasks. Metrics such as "`User Satisfaction Score`" and "`Task Success Rate`" can be developed to quantify how well the extracted relations meet user needs in real-world applications. For example, this evaluation has shown its effectiveness in related fields, which measures the quality of user experience in recommender systems [134], [135]. Although the study specifically targets recommender systems, it underscores the significance of user-centric evaluation metrics. Beyond user-centric metrics, evaluation should include: (1) cross-lingual transfer efficiency measuring performance drop across languages, (2) linguistic analysis of error patterns across typologically different languages, (3)

computational efficiency metrics for deployment scenarios, and (4) robustness measures against linguistic variations and domain shifts.

B. OPEN DIRECTIONS FROM LITERATURE ANALYSIS

In this section, we discuss several open directions identified from the papers reviewed during our literature analysis. These potential research paths are closely tied to the papers examined and may or may not remain valid as the field evolves.

The approach applied by Zou et al. [50] demonstrates that if their experiments were conducted using semi-supervised or fully supervised methods, the results could potentially improve. Similarly, Ni and Florian [68] in their work on NCL RE suggests extending their approach to additional languages, particularly those with diverse linguistic structures. Future research could also focus on integrating this method with other techniques to enhance the model's ability to handle languages with varying word orders more effectively. Beyond what is stated in the paper, a promising direction for future work could involve exploring unsupervised learning techniques to reduce dependency on bilingual dictionaries, thereby improving performance in truly low-resource languages. Specifically, self-supervised contrastive learning approaches that leverage unlabeled multilingual corpora could be combined with adversarial feature adaptation to create more robust cross-lingual representations without relying on parallel data or dictionaries.

Subburathinam et al. [65] in their work on Structure Transfer discuss future directions, noting that combining their procedure with the latest word embeddings and KG embeddings could further enhance performance. Similarly, Harting et al.'s LOREM [60] highlighted that transferring knowledge between languages from the same language family could yield more effective results when working with multilingual models. This language-family-based transfer could be formalized through meta-learning frameworks that explicitly model the typological similarities between languages, allowing for more efficient adaptation to new languages within the same family while preserving relation semantics.

Some extraction types, such as nominal relations, conjunctions in arguments, and contextual information, are not addressed in Multi^{2}OIE [61]. This opens up opportunities to explore these aspects in future studies, particularly for non-alphabetic languages that were not considered in the original paper. Moreover, PUCRJ-PUCPR-UFMG [62] highlights that many systems currently rely on multilingual BERT (mBERT), which supports 104 languages. However, these systems have not been widely evaluated on languages outside the scope of mBERT. Therefore, an interesting direction would be to assess such systems on languages that are not supported by mBERT or similar models, potentially expanding their applicability. For languages outside mBERT's coverage, techniques like vocabulary extension with language-specific tokenization, adapter-based language adaptation, and cross-lingual knowledge distillation could bridge the gap without requiring full pre-training from scratch.

In the case of LOME by Xia et al. [63], it is evident that a complete MRE system is required, as LOME only considers temporal MRE. A more comprehensive system that incorporates additional types of relational information could further advance the field. Additionally, GATE by Ahmad et al. [30] suggests that including structural information from different languages could further improve MRE performance. A unified architecture that jointly models multiple relation types (temporal, causal, spatial, etc.) could leverage shared cross-type patterns while maintaining type-specific features, potentially through a multi-task learning framework with relation-type-specific decoders operating on shared representations.

PARE by Rathore et al. [36] proposes that embeddings of entity mentions in multilingual settings could be better aligned using constrained learning techniques, which could enhance token embeddings. Constraints can be applied to label hierarchies, such as `PresidentOf` implying `CitizenOf`, since in PARE, label query vectors operate independently. They also mention that translation-based approaches during training or inference could improve the performance of mPARE. These hierarchical constraints could be formalized through logical entailment frameworks that enforce consistency across relation predictions, potentially using box embeddings or order embeddings that naturally capture hierarchical relationships between concepts across languages.

Lastly, Prompt-XRE by Hsu et al. [39] suggests that recent advances in prompt tuning could be explored with more recent PLMs to further improve MRE performance. In the case of Nag et al.'s TransRel [69], it is suggested that languages with similar structures, such as Hindi and Bengali, could be further explored. No individual models have yet been trained to target relations in these low-resource languages, and new models should be developed that either utilize existing resources or incorporate more diverse languages. Soft prompt tuning approaches that learn continuous prompt vectors specific to each language could be combined with cross-lingual alignment objectives to create language-adaptive prompts that preserve relation semantics while accommodating language-specific syntactic patterns.

C. UNTAPPED OPPORTUNITIES

There are several open research directions in the field of MRE, offering low-hanging fruits for further exploration. One critical need is the development of multilingual biomedical embeddings. While Lee et al.'s BioBERT [83] and similar models have been successful in English, there is a lack of comparable embeddings for other languages, especially within the biomedical domain. The field also lacks a universal extractor that can process relations across multiple languages and domains, which remains an essential gap to be addressed in future research. Domain-specific multilingual models could be developed through continued pre-training

of existing multilingual models on domain-specific corpora across multiple languages, with specialized objectives that capture domain-specific relation patterns while preserving cross-lingual alignment.

Lexical semantic relations form the backbone of lexical semantics and support the construction of comprehensive knowledge bases and embeddings. Our review found a significant gap in multilingual approaches specifically targeting these relations. This absence is particularly notable given the rich resources available in this domain, including cross-lingual extensions of WordNet such as BabelNet and MultiWordNet. The lack of MRE approaches focusing on lexical semantic relations represents a missed opportunity, as these fundamental relationships could serve as a bridge between languages with different structural properties. Future research should explore how lexical semantic RE techniques can be adapted for multilingual settings, potentially leveraging existing multilingual lexical resources to improve cross-lingual knowledge transfer.

Moreover, open RE systems have yet to be developed for several prominent languages such as Chinese, Japanese, and Korean, highlighting an area that remains largely unexplored. While general-purpose MRE has been studied, certain specific tasks like causal RE and document-level RE are still underdeveloped. This offers an open avenue for research that could significantly enhance the scope and utility of MRE in real-world applications. Addressing these gaps will allow MRE systems to better serve both research and practical applications, particularly in underrepresented languages. For East Asian languages, character-level and subword-level modeling approaches that account for logographic writing systems could be combined with syntactic parsing information to create more effective open RE systems that handle the unique structural properties of these languages.

a: PRACTICAL APPLICATIONS AND KNOWLEDGE GRAPH INTEGRATION

A promising direction for MRE research is the development of end-to-end pipelines that extract relations from multilingual sources to construct comprehensive KGs that transcend language boundaries. These systems could enable cross-lingual information retrieval, allowing users to query in one language and retrieve relevant content from documents in other languages. Additionally, MRE capabilities could be integrated into question answering systems to improve reasoning about relationships between entities across languages. The synergy between MRE and KGs offers opportunities for mutual enhancement through entity alignment, knowledge-enhanced extraction, and joint learning approaches that simultaneously extract relations from text and reason over structured knowledge.

VI. CONCLUSION

In this paper, we addressed the lack of a systematic literature review that comprehensively and quantitatively analyzes the landscape of MRE research. To fill this gap, we conducted a thorough review of existing works and identified open problems and challenges for future research.

Our review included 39 approaches and datasets that we meticulously annotated to capture their key characteristics. We analyzed current research phenomena and derived valuable insights. Specifically, we examined 18 research articles and 21 datasets/resources articles across six perspectives: 1) methodologies adapted, 2) number of languages explored, 3) types or domains, 4) reproducibility, 5) datasets used, and 6) evaluation metrics. We performed a comprehensive analysis based on these perspectives, categorizing all approaches and identifying further subcategories. Additionally, we compared various datasets and benchmarks proposed in the MRE literature and provided guidelines for an effective MRE dataset. We hope that this exploration of the MRE domain offers new insights for future applications and research approaches.

REFERENCES

[1] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–39, Jan. 2022, doi: 10.1145/3445965.

[2] J. Yang, S. C. Han, and J. Poon, "A survey on extraction of causal relations from natural language text," *Knowl. Inf. Syst.*, vol. 64, no. 5, pp. 1161–1186, May 2022, doi: 10.1007/s10115-022-01665-w.

[3] S. Zhou, B. Yu, A. Sun, C. Long, J. Li, and J. Sun, "A survey on neural open information extraction: Current status and future directions," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, 2022, pp. 5694–5701.

[4] H. Jiang, Q. Bao, Q. Cheng, D. Yang, L. Wang, and Y. Xiao, "Complex relation extraction: Challenges and opportunities," 2020, *arXiv:2012.04821*.

[5] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 455–465.

[6] M. Ali, M. Saleem, D. Moussallem, M. A. Sherif, and A.-C. N. Ngomo, "RELD: A knowledge graph of relation extraction datasets," in *Proc. Eur. Semantic Web Conf.*, 2023, pp. 337–353.

[7] L. Zhao, W. Alhoshan, A. Ferrari, K. J. Letsholo, M. A. Ajagbe, E.-V. Chioasca, and R. T. Batista-Navarro, "Natural language processing for requirements engineering: A systematic mapping study," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–41, Apr. 2022, doi: 10.1145/3444689.

[8] M. Ali, M. Saleem, and A.-C. N. Ngomo, "Unsupervised relation extraction using sentence encoding," in *Proc. Semantic Web, ESWC Satell. Events: Virtual Event*, 2021, pp. 136–140.

[9] P.-L. H. Cabot, S. Tedeschi, A.-C. N. Ngomo, and R. Navigli, "RED^{FM}: A filtered and multilingual relation extraction dataset," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, Jul. 2023, pp. 4326–4343. [Online]. Available: https://aclanthology.org/2023.acl-long.237

[10] L. Hennig, P. Thomas, and S. Möller, "MultiTACRED: A multilingual version of the TAC relation extraction dataset," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, Jul. 2023, pp. 3785–3801. [Online]. Available: https://aclanthology.org/2023.acl-long.210

[11] Y. Lin, Z. Liu, and M. Sun, "Neural relation extraction with multi-lingual attention," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 34–43.

[12] X. Wang, X. Han, Y. Lin, Z. Liu, and M. Sun, "Adversarial multi-lingual neural relation extraction," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1156–1166.

[13] P. Verga, D. Belanger, E. Strubell, B. Roth, and A. McCallum, "Multilingual relation extraction using compositional universal schema," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 886–896.

[14] H. Seale, B. Harris-Roxas, A. Heywood, I. Abdi, A. Mahimbo, A. Chauhan, and L. Woodland, "Speaking COVID-19: Supporting COVID-19 communication and engagement efforts with people from culturally and linguistically diverse communities," *BMC Public Health*, vol. 22, no. 1, p. 1257, Dec. 2022.

[15] *The Fukushima Daiichi Accident*, Non-serial Publications, International Atomic Energy Agency, Vienna, Austria, 2015. [Online]. Available: https://www.iaea.org/publications/10962/the-fukushima-daiichi-accident

[16] A. Seganti, K. Firląg, H. Skowronska, M. Satława, and P. Andruszkiewicz, "Multilingual entity and relation extraction dataset and model," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, Apr. 2021, pp. 1946–1955. [Online]. Available: https://aclanthology.org/2021.eacl-main.166

[17] Z. Wang, J. Yang, T. Li, J. Liu, Y. Mo, J. Bai, L. He, and Z. Li, "Multilingual entity and relation extraction from unified to language-specific training," in *Proc. Int. Conf. Electron., Comput. Commun. Technol.*, 2023, pp. 98–105.

[18] L. Kong, Y. Chu, Z. Ma, J. Zhang, L. He, and J. Chen, "MixRED: A mix-lingual relation extraction dataset," in *Proc. Joint Int. Conf. Comput. Linguistics*, May 2024, pp. 11361–11370. [Online]. Available: https://aclanthology.org/2024.lrec-main.993

[19] Y. Ma, A. Wang, and N. Okazaki, "Building a Japanese document-level relation extraction dataset assisted by cross-lingual transfer," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval. (LREC-COLING)*, 2024, pp. 2567–2579.

[20] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, and R. Xu, "A comprehensive survey on relation extraction: Recent advances and new frontiers," *ACM Comput. Surv.*, vol. 56, no. 11, pp. 1–39, Jul. 2024, doi: 10.1145/3674501.

[21] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang, "A survey of information extraction based on deep learning," *Appl. Sci.*, vol. 12, no. 19, p. 9691, Sep. 2022.

[22] H. Wang, K. Qin, R. Y. Zakari, G. Lu, and J. Yin, "Deep neural network-based relation extraction: An overview," *Neural Comput. Appl.*, vol. 34, no. 6, pp. 4781–4801, Mar. 2022, doi: 10.1007/s00521-021-06667-3.

[23] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutiérrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, 2021.

[24] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, Aug. 2007.

[25] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. 13th Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2009, pp. 147–155, doi: 10.3115/1596374.1596399. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596374.1596399

[26] I. L. Oliveira, R. Fileto, R. Speck, L. P. F. Garcia, D. Moussallem, and J. Lehmann, "Towards holistic entity linking: Survey and directions," *Inf. Syst.*, vol. 95, Jan. 2021, Art. no. 101624.

[27] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 443–460, Feb. 2015.

[28] P.-L. Huguet Cabot and R. Navigli, "REBEL: Relation extraction by end-to-end language generation," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, Punta Cana, Dominica. Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2370–2381. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.204

[29] R. Speck and A.-C. N. Ngomo, "On extracting relations using distributional semantics and a tree generalization," in *Knowledge Engineering and Knowledge Management*, C. F. Zucker, C. Ghidini, A. Napoli, and Y. Toussaint, Eds., Cham, Switzerland: Springer, 2018, pp. 424–438.

[30] W. U. Ahmad, N. Peng, and K.-W. Chang, "GATE: Graph attention transformer encoder for cross-lingual relation and event extraction," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 14, pp. 12462–12470.

[31] X. Han, T. Gao, Y. Lin, H. Peng, Y. Yang, C. Xiao, Z. Liu, P. Li, M. Sun, and J. Zhou, "More data, more relations, more context and more openness: A review and outlook for relation extraction," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics 10th Int. Joint Conf. Natural Lang. Process.*, Dec. 2020, pp. 745–758. [Online]. Available: https://aclanthology.org/2020.aacl-main.75

[32] Y. Yang, H. Xun, and H. You, "A review of relation extraction," *Literature Rev. Lang. Statist. II*, vol. 29, no. 11, pp. 30–39, 2013.

[33] Y. Chen, D. Harbecke, and L. Hennig, "Multilingual relation classification via efficient and effective prompting," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Dec. 2022, pp. 1059–1075.

[34] G. Zhou and T. He, "Using semantic similarity to enhance semi-supervised learning methods for relation extraction," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 2139–2143.

[35] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. AFNLP ACL-IJCNLP*, vol. 2, 2009, pp. 1003–1011.

[36] V. Rathore, K. Badola, P. Singla, and M., "PARE: A simple and strong baseline for monolingual and multilingual distantly supervised relation extraction," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, May 2022, pp. 340–354. [Online]. Available: https://aclanthology.org/2022.acl-short.38

[37] A. Mohamed and I. Gurevych, "Exploiting partial redundancy in unsupervised relation extraction," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technologies, Short Papers*, vol. 2, 2011, pp. 281–285.

[38] C. Yu, H. Xue, M. Wang, and L. An, "Towards an entity relation extraction framework in the cross-lingual context," *Electron. Library*, vol. 39, no. 3, pp. 411–434, Nov. 2021.

[39] C. Hsu, C. Zan, L. Ding, L. Wang, X. Wang, W. Liu, F. Lin, and W. Hu, "Prompt-learning for cross-lingual relation extraction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–9.

[40] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, "Position-aware attention and supervised data improve slot filling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 35–45.

[41] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun, "DocRED: A large-scale document-level relation extraction dataset," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 764–777.

[42] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2010, pp. 148–163.

[43] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "BioCreative V CDR task corpus: A resource for chemical disease relation extraction," *Database*, vol. 2016, Jan. 2016, Art. no. baw068.

[44] H. Khaldi, F. Benamara, G. Siegel, C. Pradel, and N. Aussenac-Gilles, "How's business going worldwide? A multilingual annotated corpus for business relation extraction," in *Proc. 13th Conf. Lang. Resour. Eval. (LREC)*, 2022, pp. 3696–3705.

[45] I. Mani, M. Verhagen, B. Wellner, C. Lee, and J. Pustejovsky, "Machine learning of temporal relations," in *Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meeting Assoc. Comput. Linguistics*, 2006, pp. 753–760.

[46] I. Vulić, S. Baker, E. M. Ponti, U. Petti, I. Leviant, K. Wing, O. Majewska, E. Bar, M. Malone, T. Poibeau, R. Reichart, and A. Korhonen, "Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity," *Comput. Linguistics*, vol. 46, no. 4, pp. 847–897, Feb. 2021.

[47] D. Gromann, "MultiLexBATS: Multilingual dataset of lexical semantic relations," in *Proc. Joint Int. Conf. Comput. Linguistics*, May 2024, pp. 11783–11793.

[48] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Jun. 2018, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[49] B. Min, Z. Jiang, M. Freedman, and R. Weischedel, "Learning transferable representation for bilingual relation extraction via convolutional neural networks," in *Proc. 8th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, Nov. 2017, pp. 674–684. [Online]. Available: https://aclanthology.org/I17-1068

[50] B. Zou, Z. Xu, Y. Hong, and G. Zhou, "Adversarial feature adaptation for cross-lingual relation classification," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 437–448.

[51] N. Taghizadeh and H. Faili, "Cross-lingual transfer learning for relation extraction using universal dependencies," *Comput. Speech Lang.*, vol. 71,

p. 12, Jan. 2022, doi: 10.1016/j.csl.2021.101265. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000711

[52] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: https://openreview.net/forum?id=H196sainb

[53] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Commun. ACM*, vol. 51, no. 12, pp. 68–74, Dec. 2008.

[54] F. S. Tsai, M. Etoh, X. Xie, W.-C. Lee, and Q. Yang, "Introduction to mobile information retrieval," *IEEE Intell. Syst.*, vol. 25, no. 1, pp. 11–15, Jan. 2010.

[55] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and Techniques*. Waltham, MA, USA: Morgan Kaufmann, 2012.

[56] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, Nov. 2001.

[57] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.

[58] B. Kitchenham, *Procedures for Performing Systematic Reviews*, vol. 33. Keele, U.K.: Keele Univ., 2004, pp. 1–26.

[59] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, Jul. 2009, Art. no. e1000097.

[60] T. Harting, S. Mesbah, and C. Lofi, "LOREM: Language-consistent open relation extraction from unstructured text," in *Proc. Web Conf.*, Apr. 2020, pp. 1830–1838, doi: 10.1145/3366423.3380252.

[61] Y. Ro, Y. Lee, and P. Kang, "Multi^{2}OIE: Multilingual open information extraction based on multi-head attention with BERT," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, Nov. 2020, pp. 1107–1117. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.99

[62] L. Pavanelli, E. T. R. Schneider, Y. B. Gumiel, T. C. Ferreira, L. F. A. de Oliveira, J. V. A. de Souza, G. P. M. Paiva, L. E. S. e Oliveira, C. M. C. Moro, E. C. Paraíso, and A. S. Pagano, "PUCRJ-PUCPR-UFMG at eHealth-KD challenge 2021: A multilingual BERT-based system for joint entity recognition and relation extraction," in *Proc. IberLEF@ SEPLN*, 2021, pp. 683–691.

[63] P. Xia, G. Qin, S. Vashishtha, Y. Chen, T. Chen, C. May, C. Harman, K. Rawlins, A. S. White, and B. Van Durme, "LOME: Large ontology multilingual extraction," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations*, Apr. 2021, pp. 149–159. [Online]. Available: https://aclanthology.org/2021.eacl-demos.19

[64] T. Caselli, I. Dini, and F. Dell'Orletta, "How about time? Probing a multilingual language model for temporal relations," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 3197–3209.

[65] A. Subburathinam, D. Lu, H. Ji, J. May, S.-F. Chang, A. Sil, and C. R. Voss, "Cross-lingual structure transfer for relation and event extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 313–325.

[66] K. Wei, L. Jin, Z. Zhang, Z. Guo, X. Li, Q. Liu, and W. Feng, "More than syntaxes: Investigating semantics to zero-shot cross-lingual relation extraction and event argument role labelling," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 5, pp. 1–21, May 2024, doi: 10.1145/3582261.

[67] M. Faruqui and S. Kumar, "Multilingual open relation extraction using cross-lingual projection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2015, pp. 1351–1356.

[68] J. Ni and R. Florian, "Neural cross-lingual relation extraction based on bilingual word embedding mapping," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Nov. 2019, pp. 399–409. [Online]. Available: https://aclanthology.org/D19-1038

[69] A. Nag, B. Samanta, A. Mukherjee, N. Ganguly, and S. Chakrabarti, "Transfer learning for low-resource multilingual relation classification," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 2, pp. 1–24, Feb. 2023, doi: 10.1145/3554734.

[70] S. Zhang, K. Duh, and B. Van Durme, "MT/IE: Cross-lingual open information extraction with neural sequence-to-sequence models," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, Short Papers*, vol. 2, 2017, pp. 64–70.

[71] S. Zhang, K. Duh, and B. V. Durme, "Selective decoding for cross-lingual open information extraction," in *Proc. 8th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2017, pp. 832–842.

[72] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A survey on open information extraction," in *Proc. 27th Int. Conf. Comput. Linguistics*, Aug. 2018, pp. 3866–3878. [Online]. Available: https://aclanthology.org/C18-1326

[73] G. Stanovsky, J. Michael, L. Zettlemoyer, and I. Dagan, "Supervised open information extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 885–895.

[74] V. Shwartz and I. Dagan, "Still a pain in the neck: Evaluating text representations on lexical composition," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 403–419, Nov. 2019.

[75] H. Ji, J. Nothman, and B. Hachey, "Overview of TAC-KBP2014 entity discovery and linking tasks," in *Proc. Text Anal. Conf. (TAC2014)*, 2014, pp. 1333–1339.

[76] Q. Wei, Z. Ji, Y. Si, J. Du, J. Wang, F. Tiryaki, S. Wu, C. Tao, K. Roberts, and H. Xu, "Relation extraction from clinical narratives using pre-trained language models," in *Proc. AMIA Annu. Symp.*, Mar. 2019, pp. 1236–1245.

[77] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 5, pp. 507–513, Sep. 2010, doi: 10.1136/jamia.2009.001560.

[78] C. Kedzie, K. McKeown, and F. Diaz, "Predicting salient updates for disaster summarization," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2015, pp. 1608–1617.

[79] B. Zhou, D. Khashabi, N. Qiang, and D. Roth, "'Going on a vacation' takes longer than 'going for a walk': A study of temporal commonsense understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Nov. 2019, pp. 3363–3369. [Online]. Available: https://aclanthology.org/D19-1332

[80] Q. Ning, Z. Feng, H. Wu, and D. Roth, "Joint reasoning for temporal and causal relations," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, Jul. 2018, pp. 2278–2288. [Online]. Available: https://aclanthology.org/P18-1212

[81] C. Blaschke, M. A. Andrade-Navarro, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: Protein–protein interactions," in *Proc. ISMB*, 1999, pp. 60–67.

[82] Z. Wei, J. Su, Y. Wang, Y. Tian, and Y. Chang, "A novel cascade binary tagging framework for relational triple extraction," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 1476–1488. [Online]. Available: https://aclanthology.org/2020.acl-main.136

[83] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[84] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," 2020, *arXiv:2010.02559*.

[85] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[86] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–41, Jun. 2020.

[87] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.* San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 260–270. [Online]. Available: https://aclanthology.org/N16-1030

[88] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, *arXiv:1603.01360*.

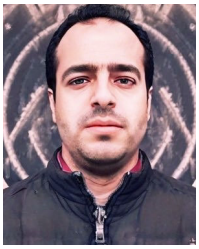
[89] D. Zhang and D. Wang, "Relation classification via recurrent neural network," 2015, *arXiv:1508.01006*.

[90] C. Alt, M. Hübner, and L. Hennig, "Fine-tuning pre-trained transformer language models to distantly supervised relation extraction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1388–1398.
[91] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[92] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.
[93] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Jul. 2020, pp. 6282–6293. [Online]. Available: https://aclanthology.org/2020.acl-main.560/
[94] C. Walker and L. D. Consortium, *ACE 2005 Multilingual Training Corpus* (LDC Corpora). Philadelphia, PA, USA: Linguistic Data Consortium, 2005. [Online]. Available: https://books.google.de/books?id=SbjjuQEACAAJ
[95] D. I. Adelani, "MasakhaNER: Named entity recognition for African languages," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1116–1131, Jan. 2021.
[96] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2019, pp. 4996–5001. [Online]. Available: https://aclanthology.org/P19-1493
[97] B. Plank, D. Hovy, and A. Søgaard, "Learning part-of-speech taggers with inter-annotator agreement loss," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 742–751.
[98] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. COLING 25th Int. Conf. Comput. linguistics, Tech. Papers*, 2014, pp. 2335–2344.
[99] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: Evaluating cross-lingual sentence representations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2475–2485. [Online]. Available: https://aclanthology.org/D18-1269
[100] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[101] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 587–604, Dec. 2018. [Online]. Available: https://aclanthology.org/Q18-1041
[102] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, 2009.
[103] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. (NLP)*, Nov. 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446
[104] E. Bassignana, F. Ginter, S. Pyysalo, R. van der Goot, and B. Plank, "Multi-CrossRE a multi-lingual multi-domain dataset for relation extraction," in *Proc. 24th Nordic Conf. Comput. Linguistics (NoDaLiDa)*. Faroe Islands: University Tartu Library, May 2023, pp. 80–85. [Online]. Available: https://aclanthology.org/2023.nodalida-1.9
[105] E. Bassignana and B. Plank, "CrossRE: A cross-domain dataset for relation extraction," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3592–3604. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.263
[106] A. P.-M. Suilan, E.-Velarde, Y. Gutierrez, Y. Almeida-Cruz, A. Montoyo, and R. Muñoz, "Overview of the eHealth knowledge discovery challenge at iberLEF 2021," *Procesamiento del Lenguaje Natural*, vol. 67, pp. 233–242, Jan. 2021. [Online]. Available: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6392
[107] J. Zhan and H. Zhao, "Span model for open information extraction on accurate corpus," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 5, pp. 9523–9530, doi: 10.1609/aaai.v34i05.6497.
[108] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, and L. Ferro, *The Timebank Corpus*. Lancaster, U.K.: Corpus linguistics, 2003, p. 40.
[109] T. Cassidy, B. McDowell, N. Chambers, and S. Bethard, "An annotation framework for dense event ordering," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, Jun. 2014, pp. 501–506. [Online]. Available: https://aclanthology.org/P14-2082
[110] R. Saurı and T. Badia, "Spanish timebank," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep. LDC2012T12, 2012. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2012T12
[111] T. Caselli, V. Bartalesi, R. Sprugnoli, E. Pianta, and I. Prodanof, "Annotating events, temporal expressions and relations in Italian: The it-timeml experience for the ita-TimeBank," in *Proc. 5th Linguistic Annotation Workshop*, 2011, pp. 143–151.
[112] A. Bittar, P. Amsili, P. Denis, and L. Danlos, "French TimeBank: An ISO-TimeML annotated reference corpus," in *Proc. 49th Annu. meeting Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2011, pp. 130–134.
[113] A. Gladkova, A. Drozd, and S. Matsuoka, "Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 8–15.
[114] B. C. L. Ferreira, C. Silva, and H. G. Oliveira, "Towards automated evaluation of knowledge encoded in large language models," in *Proc. Workshop Deep Learn. Linked Data (DLnLD) LREC-COLING*, May 2024, pp. 76–85.
[115] H. Hassan, "Achieving human parity on automatic Chinese to english news translation," 2018, *arXiv:1803.05567*.
[116] A. Bhartiya, K. Badola, and M., "DiS-ReX: A multilingual dataset for distantly supervised relation extraction," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, Dublin, Ireland, 2022, pp. 849–863, doi: 10.18653/v1/2022.acl-short.95.
[117] V. Lai, H. Man, L. Ngo, F. Dernoncourt, and T. Nguyen, "Multilingual SubEvent relation extraction: A novel dataset and structure induction method," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, Abu Dhabi, United Arab Emirates, 2022, pp. 5559–5570, doi: 10.18653/v1/2022.findings-emnlp.407.
[118] A. Köksal and A. Özgür, "The RELX dataset and matching the multilingual blanks for cross-lingual relation classification," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, Nov. 2020, pp. 340–350, doi: 10.18653/v1/2020.findings-emnlp.32.
[119] A. Plum, T. Ranasinghe, and C. Purschke, "Guided distant supervision for multilingual relation extraction data: Adapting to a new language," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval. (LREC-COLING)*, 2024, pp. 7982–7992.
[120] A. Ananthram, E. Allaway, and K. McKeown, "Event-guided denoising for multilingual relation learning," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, Dec. 2020, pp. 1505–1512. [Online]. Available: https://aclanthology.org/2020.coling-main.131
[121] M. Abdou, C. Sas, R. Aralikatte, I. Augenstein, and A. Søgaard, "X-WikiRE: A large, multilingual resource for relation extraction as machine comprehension," in *Proc. 2nd Workshop Deep Learn. Approaches Low-Resource NLP (DeepLo)*, Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 265–274. [Online]. Available: https://aclanthology.org/D19-6130
[122] B. Xu, Q. Wang, Y. Lyu, D. Dai, Y. Zhang, and Z. Mao, "S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2023, pp. 8186–8207.
[123] D. B. Claro, M. Souza, C. C. Xavier, and L. Oliveira, "Multilingual open information extraction: Challenges and opportunities," *Information*, vol. 10, no. 7, p. 228, Jul. 2019.
[124] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," 2017, *arXiv:1707.07328*.
[125] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," 2021, *arXiv:2104.08821*.
[126] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *Proc. Int. Conf. Mach. Learn.*, vol. 1, 2020, pp. 4411–4421.
[127] Q. Wang, Y. Mao, J. Wang, H. Yu, S. Nie, S. Wang, F. Feng, L. Huang, X. Quan, and Z. Xu, "APrompt: Attention prompt tuning for efficient adaptation of pre-trained language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 9147–9160.
[128] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.
[129] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17283–17297.
[130] R. Takanobu, T. Zhang, J. Liu, and M. Huang, "A hierarchical framework for relation extraction with reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 7072–7079.

[131] X. Han, P. Yu, Z. Liu, M. Sun, and P. Li, "Hierarchical relation extraction with coarse-to-fine grained attention," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2236–2245.
[132] T.-J. Fu, P.-H. Li, and W.-Y. Ma, "GraphRel: Modeling text as relational graphs for joint entity and relation extraction," in *Proc. 57th Annu. Meeting Assoc. Comput. linguistics*, 2019, pp. 1409–1418.
[133] S. Wadhwa, S. Amir, and B. Wallace, "Revisiting relation extraction in the era of large language models," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2023, p. 15566.
[134] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proc. 5th ACM Conf. Recommender Syst.*, Oct. 2011, pp. 157–164.
[135] D. Kotkov, A. Medlar, T. Kask, and D. Glowacka, "The dark matter of serendipity in recommender systems," in *Proc. ACM SIGIR Conf. Human Inf. Interact. Retr.*, Mar. 2024, pp. 108–118.

MANZOOR ALI received the M.Sc. degree in software engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2016. He is currently pursuing the Ph.D. degree with the Data Science Group, Paderborn University, Germany. His research interests include natural language processing, information extraction—with a particular focus on relation extraction, semantic web technologies, generative artificial intelligence, and machine learning.

RENÉ SPECK received the M.Sc. degree in computer science from AKSW, Leipzig University, Germany, in 2014. He is currently pursuing the Ph.D. degree with the Data Science Group (DICE), Paderborn University, Germany. Since then, he has developed several widely used tools and frameworks. His research interests include research topics related to knowledge graphs and semantic web technologies, particularly knowledge graph population, relation extraction, and machine learning.



HAMADA M. ZAHERA received the Ph.D. degree in computer science from Paderborn University. He is a Postdoctoral Researcher with the Data Science Group (DICE). Currently, he leads the NLP unit at DICE, where he focuses on developing novel methods for NLP and large language models. Additionally, he co-supervises Ph.D. students on various topics, including question answering, relation extraction, and entity summarization.

MUHAMMAD SALEEM received the Ph.D. degree in computer science from AKSW, Leipzig University. He is currently a Unit Leader of Data Storage and Querying (DSQ) and a Senior Researcher with Paderborn University. His research interests include machine learning and AI techniques for SPARQL query processing and benchmarking, federated SPARQL query optimization and source selection, linked data summaries, and top-k query processing.

DIEGO MOUSSALLEM received the Ph.D. degree in computer science and has a background in NLP, working on tasks, such as entity linking, machine translation, and natural language generation. He has contributed to the scientific community as the Portuguese Chapter of DBpedia and a Google Summer of Code Mentor. He is currently the Head of Data Science of Jusbrasil and was previously a Lead Data Scientist with Globo, an NLP team lead at Paderborn University.

AXEL-CYRILLE NGONGA NGOMO is currently the Data Science (DICE) Chair of the Computer Science Department, Paderborn University. He has co-authored more than 200 reviewed publications and has developed several widely used frameworks. His research interests include areas around knowledge graphs and semantic web technologies, especially link discovery, federated queries, machine learning, and natural-language processing.