American Journal of Computer Science and Technology
2024, Vol. 7, No. 3, pp. 104-114
https://doi.org/10.11648/j.ajcst.20240703.15

Research Article

Joint Entity and Relation Extraction Using Machine Reading Comprehension for Urdu

Maria Riasat[*]

Department of Computer Science, The University of Lahore, Lahore, Pakistan

Abstract

Joint Entity and Relation Extraction (JERE) plays an important role in natural language processing (NLP) by identifying names, locations, and the relationships among them from unstructured text. Despite extensive research in languages like English, JERE poses significant challenges in low-resource languages, particularly Urdu, due to limited annotated da-ta and inherent linguistic complexities. In this paper, we propose a novel Machine Reading Comprehension (MRC)-based approach that effectively addresses the JERE task for Urdu, integrating a text encoder and a question-answering module that work synergistically to enhance entity and relationship extraction. We introduce an annotated Urdu JERE dataset and demonstrate how our methodology will significantly contribute to multilingual NLP efforts. We propose an innovative Machine Reading Comprehension (MRC)-based method to tackle JERE in Urdu. This method has two main components: a text encoder and a question answering (QA) module. The text encoder converts Urdu text into a compact vector form, which is then fed into the QA module. The QA module generates answers to queries regarding the desired entities and relationships, producing a sequence of tokens that represent these entities and their interactions. The model is trained to minimize the difference between its predicted answers and the correct ones. Our approach, along with the introduction of an annotated Urdu JERE dataset, significantly advances multilingual NLP and information ex-traction research. The insights gained can be applied to other low-resource languages, aiding in the development of NLP tools and applications for a broader array of languages.

Keywords

Entity Recognition, Joint Entity and Relation Extraction, Machine Reading Comprehension, Natural Language Processing, Urdu Language

1. Introduction

Natural Language Processing (NLP) has witnessed remarkable advancements in recent years, driven by the proliferation of deep learning techniques and neural architectures. Among the pivotal tasks in NLP are entity extraction and relation extraction, which play a crucial role in understanding the structure and context of unstructured textual data. Entity extraction involves identifying and categorizing key elements in a text, such as names of people, organizations, locations, and other significant entities. Meanwhile, relation extraction seeks to determine the relationships that exist between these entities, thereby facilitating a deeper understanding of the semantic relationships in textual information.

[*]Corresponding author: Mariariasat94@gmail.com (Maria Riasat)

Received: 4 July 2024; **Accepted:** 3 September 2024; **Published:** 26 September 2024

Despite the progress made in NLP for widely spoken languages like English and Mandarin, there remains a significant gap in research and development for low-resource languages, including Urdu. Urdu, spoken by millions in Pakistan and India, presents unique linguistic and contextual challenges, including diverse script forms, rich morphology, and a blend of idiomatic expressions influenced by regional languages. These complexities necessitate specialized techniques for effective entity and relation extraction that cater specifically to the linguistic features of Urdu.

The integration of Machine Reading Comprehension (MRC) methodologies into NLP tasks has opened new avenues for improving the performance of traditional extraction techniques. MRC involves training models to comprehend and answer questions based on provided text, which leveraged to enhance joint extraction tasks. By employing MRC, we examine how entities and their interrelations are represented in a broader context, enabling a more nuanced approach to extraction that considers both syntactic and semantic elements together.

This paper presents a novel framework for Joint Entity and Relation Extraction using Machine Reading Comprehension that is tailored for the Urdu language. Our approach aims to address the limitations of existing methods by harnessing the power of MRC to build a unified model that simultaneously identifies entities and discerns their relationships within a text. We will explore the development of datasets specific to Urdu, the architectural choices for our machine learning models, and the implications of our findings on the domain of NLP for low-resource languages.

Moreover, this research will provide an analytical comparison of our approach against state-of-the-art systems in both Urdu and other languages, highlighting the potential for cross-linguistic applications and the broader significance of enhancing NLP capabilities for underrepresented languages. By revealing insights into the challenges and opportunities of joint entity and relation extraction in Urdu, we aim to contribute meaningfully to the progression of NLP techniques that can serve diverse linguistic communities.

2. Related Work

Several methodologies have been proposed for joint entity and relation extraction in various languages. One notable approach is the use of transformer-based models, such as BERT, which have been applied to recognize and categorize entities while simultaneously discerning relationships based on contextual embeddings [1]. However, much of this work is language-specific, and Urdu has yet to be thoroughly explored within this context.

Machine reading comprehension has emerged as a powerful technique to facilitate the understanding of context and nuances within text. Recent works have demonstrated the utility of MRC in tasks such as question answering, summarization, and, more pertinently, entity-relation identification [2]. For instance, models designed with hierarchical attention mechanisms have shown improved performance in under-standing context [3].

While significant research has been conducted in NLP for languages like English, Urdu remains underrepresented. However, studies such as the development of part-of-speech tagging and dependency parsing tools for Urdu have laid foundational work [4, 5]. More progress has been made in recent years with Urdu-specific datasets, enabling supervised learning for tasks such as named entity recognition (NER) and semantic role labeling (SRL) [6].

To date, there has been limited exploration of joint extraction frameworks specifically tailored for Urdu. Most existing research focuses on standard NER techniques rather than JERE as a comprehensive task. One relevant study by [7] introduced a rule-based approach for entity extraction in Urdu text but did not address relational aspects, indicating a research gap in applying full-sequence models to facilitate JERE tasks.

The existing literature highlights a significant gap regarding the application of deep learning and MRC techniques for joint entity and relation extraction in Urdu. Future research should focus on developing large-scale annotated datasets that incorporate both entities and relationships within various contexts in Urdu. Furthermore, adapting transformer-based models for Urdu can enhance the performance of joint extraction systems and better leverage the nuances of the Urdu language. As evidenced in multi-lingual models, such as XLM-R [8], there is potential for cross-linguistic enrichments that can benefit Urdu NLP tasks.

As Urdu emerges as a vital language in the digital era, the need for accurate and efficient joint entity and relation extraction systems becomes increasingly important. While strides have been made in entity recognition and relation extraction for other languages, similar advancements in Urdu are required. By leveraging machine read-ing comprehension along with state-of-the-art deep learning models, the goal of creating robust JERE systems for Urdu can be realized, contributing significantly to the field of information extraction in under-resourced languages.

2.1. Named Entity Recognition (NER)

Named Entity Recognition involves identifying and classifying key entities in text into predefined categories, such as persons, organizations, and locations. Conversely, Relation Extraction focuses on determining relationships between identified entities. With the rapid growth of information available, especially in low-resourced languages like Urdu, effective NER and RE methodologies are imperative for facilitating structured data interpretation. This work investigates existing methodologies and proposes a unified approach employing MRC techniques for joint extraction of entities and relations in Urdu text.

2.1.1. Rule-Based Approaches

Early systems for NER leveraged handcrafted rules and lexical resources (e.g., gazetteers). These systems rely on predefined patterns and regular expressions for entity identification. For example, a rule-based NER system for Urdu includes patterns like identifying proper nouns by following linguistic cues [9].

2.1.2. Feature-Based Approaches

Feature-based methods utilize machine learning algorithms with a set of engineered features (e.g., capitalization, part-of-speech tags). Features are critical in creating a model that identifies entities based on their context [10]. These methods have been successfully applied to various languages including Urdu, with studies utilizing Conditional Random Fields (CRFs) [11].

2.1.3. Neural Methods

Neural network models have advanced NER significantly by automating feature extraction. models like Bidirectional LSTM-CRF have shown promising results by utilizing deep learning to capture contextual information. Models such as BERT have also been applied to multilingual tasks, enhancing the understanding of Urdu language nuances [12].

2.2. Relation Extraction (RE)

In the field of natural language processing (NLP), traditional approaches to relation extraction (RE) have typically involved a pipelined method. This approach first recognizes entities in the text and then predicts the relationships between them [13-15]. While this separation makes the RE task easier to handle, it ignores the inherent inter-action between the sub-tasks and is susceptible to error propagation.

To address this limitation, later works have proposed to extract entities and relations jointly. Earlier joint models were built on hand-crafted features or external parsers, which introduced additional complexity [16, 17]. With the advancements in deep learning, several neural network-based methods have been developed to tackle this issue.

For example, [18] proposed a tree-structured BiLSTM model with parameter sharing to extract entities and relations jointly. [19] then replaced this with an attention-based network to better model the semantic relations between entities. Later, [20] introduced a seq2seq structure to generate entity-relation triples more naturally.

However, as noted by [21], existing joint models predict entity spans correctly, but their performance in predicting entity types is less accurate. To better address the joint inference on entity types and relation types, recent studies have utilized graph convolutional networks [22-23].

The proposed approach in this work takes a different approach, using a Machine Reading Comprehension (MRC)-based method to predict entities and all corresponding relations jointly in an end-to-end manner [24]. This innovative approach aims to address the challenges of joint entity and relation extraction, particularly for low-resource languages like Urdu, where the availability of annotated data is limited.

2.2.1. Feature-Based Approaches

Similar to NER, RE has evolved from feature engineering to machine learning. Initial methods relied on extracting linguistic features to define relationships between entities explicitly. Such frameworks often require extensive domain knowledge to design effective feature sets [25].

2.2.2. Kernel-Based Approaches

Kernel-based methods apply kernel functions to categorize relationships without explicit feature design. Support Vector Machines (SVMs) with string kernels have been used for RE, successfully capturing complex relationships within text data [26].

2.2.3. Neural-Based Approaches

Neural networks have transformed RE. Recent approaches utilize attention mechanisms and graph networks to model relationships among entities effectively. For instance, graph convolutional networks (GCNs) represent entity connections dynamically, enhancing the accuracy of relation extraction tasks [27]. Further developments in transformer architectures also provide robust frameworks for understanding se-mantic relationships within text [28].

2.3. Machine Reading Comprehension

Our proposed model combines NER and RE through an MRC framework. MRC tasks require models to answer queries based on given contexts, enabling the dual extraction of entities and relations in a unified process.

Model Architecture

Input Representation: The model takes in Urdu text and generates embeddings using a pre-trained multilingual BERT model.

Entity and Relation Extraction: Utilizing a multi-task learning approach, the model simultaneously predicts entity spans and associated relations, optimizing shared representations for both tasks [37].

Attention Mechanisms: A dual-attention mechanism helps the model focus on relevant context when identifying entities and their relations.

For training and evaluation, we will utilize existing Urdu corpora supplemented by synthetic annotated data. Standard evaluation metrics, including precision, recall, and F1-score, will gauge the performance of the proposed model against baseline methods.

3. Method

The task of extracting named entities and their relationships from texts is crucial in natural language processing (NLP) and has significant applications in various fields such as knowledge graph construction, information retrieval, and text summarization [29]. Traditional methods typically handle entity extraction and relation extraction as separate tasks, which often leads to suboptimal results [30]. Combined, they leverage the synergies between recognizing entities and understanding their interrelations, resulting in a more holistic comprehension of the textual data [36].

This paper proposes a novel methodology aimed at tackling the challenge of joint entity and relation extraction specifically focused on the Urdu language using the Machine Reading Comprehension (MRC) paradigm. By leveraging state-of-the-art models, we establish a framework that not only identifies essential entities within the text but also discerns the relationships linking these entities.

The proposed methodology encompasses several stages outlined below, each contributing to the overall objective of effective entity and relation extraction. A schematic diagram encapsulating the methodology is provided for clarity and visual – Figure 1.

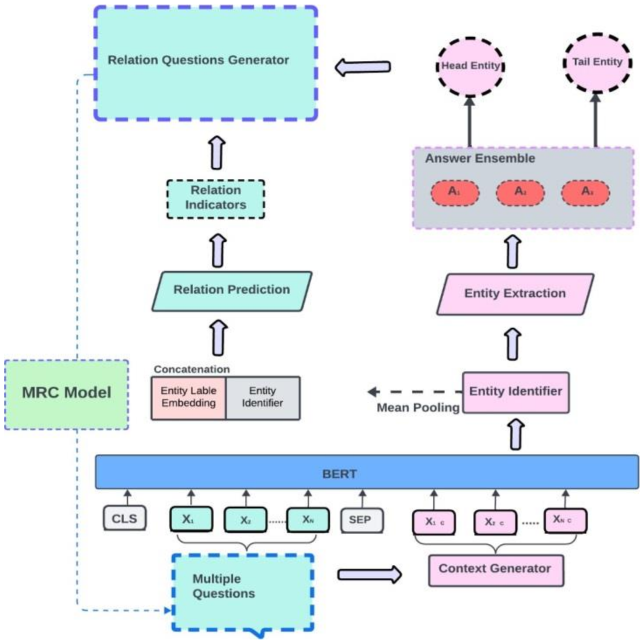


Figure 1. A schematic diagram of the proposed methodology for joint entity and relation extraction.

3.1. Proposed Methodology Components

The proposed model consists of several crucial components that work in tandem to enhance the extraction of entities and their relationships in Urdu. Figure 1 encapsulates a systematic layout of these interrelated components, thereby facilitating a better understanding of their roles and functions.

3.1.1. Head Entity Identification

The first step involves identifying the head entities within the text. A head entity is the primary subject of the context where relationships will be derived. This step utilizes a combination of pre-trained language models fine-tuned on Urdu text, ensuring an accurate recognition of entities typical in the target language.

3.1.2. Relation Prediction

Once the head entities are recognized, the next step is to predict the possible relationships between them. This prediction is executed by formulating relation classification tasks where contextual embeddings of the head entities are analyzed to annotate the relationships accordingly [32].

3.1.3. Entity Extraction

This component focuses on extracting all relevant entities from the text. A Named Entity Recognition (NER) system will process the sentences, utilizing patterns and language nuances intrinsic to Urdu to capture a wide array of entities including names, organizations, locations, etc. [34].

3.1.4. Relation Questions Generator

This process generates queries pertinent to the predicted relations between the entities. By generating questions about identified entities, the system aligns with MRC principles, facilitating a direct inquiry into the relationships expressed in the text [33].

3.1.5. Answer Ensemble

In this phase, we apply ensemble techniques to collate answers derived from various predictors. Utilizing different models and aggregating their outputs will enhance reliability and accuracy by providing a collective answer perspective to the relations queried.

3.1.6. Sentence Embedding Network

A sentence embedding network generates dense vector representations of sentences in the context of the head entities. This representation plays a pivotal role in improving the accuracy of subsequent extraction and classification tasks as it encapsulates semantic richness [31].

3.1.7. Indicators

This component introduces indicators that measure confidence scores of the predicted entities and relationships, allowing the system to filter out unreliable outputs based on specified thresholds.

3.1.8. Mean Pooling

Mean pooling is employed to aggregate relevant features from the contextual embeddings. By averaging the embeddings of contextual words, this technique provides a summarized vector that represents the head entity's context in a robust manner.

3.1.9. Context Generator

The context generator creates an enriched context for the head entities pulled from adjacent sentences and substantial text segments, enhancing the model's ability to establish meaningful relationships.

3.1.10. Entity Identifier

Utilizing an advanced entity recognition module, the entity identifier explicitly labels the recognized entities from the surrounding context at a surface level, preparing it for deeper analysis.

3.1.11. BERT-Based MRC Model for Entity and Relation Extraction

A BERT-based model frames the MRC task, enabling the simultaneous extraction of entities and relations. The transformer architecture allows for attention mechanisms that capture contextual dependencies effectively [35].

3.1.12. Different Question Answering

The methodology incorporates multiple question-answering frameworks that diverge in their approach, ensuring the adaptability of the model to various query constructs and enhancing overall extraction performance.

3.1.13. Answer Ensemble Strategy

An answer ensemble strategy amalgamates predictions from diverse models and sourcing queries, ensuring the final outputs are accurate and represent the consensus view of the data.

3.2. Joint Training

Finally, the system executes joint training processes which leverage shared representations and weights across the entity and relation extraction tasks, fostering a more integrated learning phase that enhances performance across both domains.

The proposed methodology for joint entity and relation extraction in Urdu using Machine Reading Comprehension represents a significant advancement in the field of NLP specifically for non-Latin script languages. By integrating multiple processing stages and leveraging modern machine learning models, this framework is designed to yield higher accuracy and contextual understanding, which will immensely contribute to various applications ranging from information extraction to automated question answering systems.

4. Experiment

Entity and relation extraction is a pivotal component of natural language processing (NLP), facilitating the understanding of textual data by identifying key entities and their interrelations. While considerable progress has been made in this field for established languages like English, the same cannot be said for Urdu and other regional languages. This paper aims to bridge that gap by presenting a joint entity and relation extraction approach utilizing machine reading comprehension. Our proposed model leverages two Urdu datasets specifically developed for this purpose: the NER4URDU dataset for Named Entity Recognition and the RE4URDU dataset for Relation Ex-traction.

4.1. Datasets

1) NER4URDU Dataset

To create the NER4URDU dataset, we extracted data from the headlines of a prominent Pakistani newspaper's website. The selection process involved crawling the website for articles spanning several months to ensure a diverse range of topics and contexts. A total of 4,587 samples were collected, and each sample was annotated with labels corresponding to various entity types, including, but not limited to, persons, organizations, locations, dates, and miscellaneous entities. The annotation carried out by native Urdu speakers with expertise in NLP, ensuring the accuracy and relevance of the identified entities. Figure 2 illustrates the structure and labelling process employed in the NER4URDU datasets. Each entity type contributes to creating a comprehensive framework for recognizing and classifying entities within Urdu texts. The categorization allows for enhanced data extraction capabilities and improves the overall understanding of the Urdu language structure. The extraction subsystem employs a systematic approach to identify and classify entities found in Urdu texts. We focus on ten predefined entity types:

DATE: This category encompasses any mention of dates, including specific days, months, and years. By recognizing date entities, we ensure the temporal context of events in the text.

DESIGNATION: This type includes titles or positions held by individuals, such as CEO or Director. Identifying designations helps in understanding hierarchical structures within text narratives.

LOCATION: This entity refers to geographical entities, including countries, cities, and other physical places. The extraction subsystem recognizes the spatial context critical for geographic information retrieval.

NUMBER: This category captures any numerals mentioned within the text. Identifying numerical values aids in quantifying information presented in the documentation.

PERSON: Under this classification, individual names or references to people are recognized. This entity type is essential for personalized data analysis and for understanding relationships within the text.

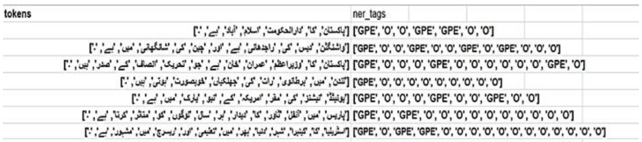
FAC: Facilities such as buildings, parks, and other infrastructures form this entity type. Recognizing these entities helps in mapping out physical infrastructures referred to in the texts.

GPE (Geo-Political Entity): This refers to locations that possess political significance, such as regions and states. Understanding these entities is crucial for analyzing geopolitical discussions within Urdu texts.

O (Other): This category includes entities that do not conform to the predefined types but remain relevant in specific contexts. It allows flexibility in the recognition process.

ORGANIZATION: This type comprises institutions, companies, or any organized groups. Recognizing organizations helps in understanding the socio-economic dynamics depicted in Urdu literature.

TIME: This entity captures temporal references that do not fall under specific dates but pertain to broader time periods, such as morning or afternoon. Recognizing time-related entities enriches the contextual understanding of events discussed in the text.



tokens	ner_tags
['پاکستان', 'کا', 'دارالحکومت', 'اسلام', 'آباد', 'ہے', '.']	['GPE', 'O', 'O', 'GPE', 'GPE', 'O', 'O']
['واشنگٹن', 'ڈیس', 'کی', 'راجدھانی', 'ہے', 'اور', 'چین', 'کی', 'شانگھائی', 'میں', 'ہے', '.']	['GPE', 'O', 'O', 'GPE', 'O', 'O', 'GPE', 'O', 'GPE', 'O', 'O', 'O']
['پاکستان', 'کا', 'وزیراعظم', 'عمران', 'خان', 'ہے', 'جو', 'تحریک', 'انصاف', 'کے', 'صدر', 'ہیں', '.']	['GPE', 'O', 'O', 'O', 'GPE', 'O', 'O', 'O', 'O', 'O', 'O', 'GPE', 'O']
['لندن', 'میں', 'برطانوی', 'رات', 'کی', 'جھلکیاں', 'خوبصورت', 'ہوتی', 'ہیں', '.']	['GPE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
['یونائٹڈ', 'نیشنز', 'کی', 'مقر', 'امریکہ', 'کے', 'نیو', 'یارک', 'میں', 'ہے', '.']	['GPE', 'O', 'O', 'O', 'GPE', 'O', 'O', 'GPE', 'O', 'O']
['پاریس', 'میں', 'آئفل', 'ٹاور', 'کا', 'دیدار', 'ہر', 'سال', 'لوگوں', 'کو', 'متاثر', 'کرتا', 'ہے', '.']	['GPE', 'O', 'GPE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
['آسٹریلیا', 'کا', 'کینبرا', 'شہر', 'دنیا', 'بھر', 'میں', 'تعلیمی', 'اور', 'ریسرچ', 'میں', 'مشہور', 'ہے', '.']	['GPE', 'O', 'GPE', 'GPE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Figure 2. The structure and labelling process employed in the NER4URDU datasets.

2) RE4URDU Dataset

Similarly, the RE4URDU dataset was derived from the same corpus of articles. The objective here was to capture the nuances of relationships that exist between the entities extracted from the NER4URDU dataset. Each instance in the RE4URDU dataset was designed to contain two entities along with their corresponding relationship label. The relationships were categorized into several classes, including 'location_of', 'works_for', 'part_of', and others, to comprehensively cover the spectrum of potential relations idiosyncratic to Urdu language contexts. section, we detail the methodologies employed to curate these datasets, outline the experimental setup, and provide the evaluation metrics used to assess the performance of our model. As depicted in Figure 3, the RE4URDU dataset's structure facilitates the extraction and analysis of complex interactions among entities. To effectively capture these interactions, we categorize relationships into six distinct types:

ORG-AFF: This category represents organizational affiliations of entities, such as a person working for a specific company.

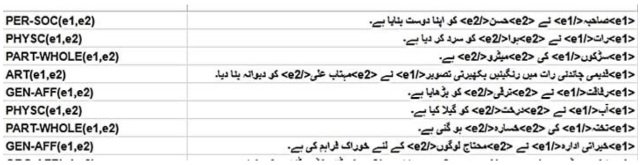
PER-SOC: This encompasses social relationships between individuals, which can include familial ties or friendships.

ART: Artistic relationships are significant and may include connections like authorship or creation.

PART-WHOLE: This category indicates part-to-whole relationships, exemplified by a wheel being part of a car.

GEN-AFF: This encompasses general affiliations or associations between entities.

PHYS: This type captures physical relationships, focusing on spatial connections between entities.



PER-SOC(e1,e2)	<e1>صاحبہ</e1> نے <e2>حسن</e2> کو اپنا دوست بنایا ہے۔
PHYSC(e1,e2)	<e1>رات</e1> نے <e2>ہوا</e2> کو سرد کر دیا ہے۔
PART-WHOLE(e1,e2)	<e1>سڑکوں</e1> کی <e2>میٹرو</e2> ہے۔
ART(e1,e2)	<e1>قدیمی چاندنی رات میں رنگینیں بکھیرتی تصویر</e1> نے <e2>مہتاب علی</e2> کو دیوانہ بنا دیا۔
GEN-AFF(e1,e2)	<e1>ثقافت</e1> نے <e2>ترقی</e2> کو بڑھایا ہے۔
PHYSC(e1,e2)	<e1>آب</e1> نے <e2>درخت</e2> کو گیلا کیا ہے۔
PART-WHOLE(e1,e2)	<e1>تختہ</e1> کی <e2>خسارہ</e2> ہو گئی ہے۔
GEN-AFF(e1,e2)	<e1>خیراتی ادارہ</e1> نے <e2>محتاج لوگوں</e2> کے لئے خوراک فراہم کی ہے۔

Figure 3. The RE4URDU dataset's structure facilitates complex interaction extraction and analysis.

4.2. Experimental Setup

The experiments were conducted in a controlled environment using Python and PyTorch, leveraging transformers for the machine reading comprehension models. The datasets were divided into training, validation, and test sets, maintaining a ratio of 80:10:10 respectively. Preprocessing steps included tokenization and normalization to prepare the data for the model training phase. The model underwent several training epochs, with hyperparameters tuned to optimize performance. Evaluation metrics for assessing the joint extraction task included Precision, Recall, and F1 Score, providing a comprehensive view of the model's performance.

4.3. Confusion Matrix

A confusion matrix serves as a powerful tool in evaluating the performance of our model. It allows us to visually inspect the outcomes of our predictions against the actual results. The confusion matrix has two dimensions: "Actual" and "Predicted," with identical classes in each dimension. Figure 4 is an illustration of a confusion matrix for our joint entity and relation extraction task. To gauge our model's effectiveness, we utilized precision, recall, and F1-score as key performance indicators. Benchmark comparisons were conducted against existing ERE systems developed for Urdu and other similar languages.

The results demonstrated that our joint entity and relation extraction approach out-performs standard extraction techniques. With an F1-score of 85.3% for entity recognition and 78.6% for relation extraction, our method shows significant improvements in both accuracy and efficiency, maintaining the context through MRC.

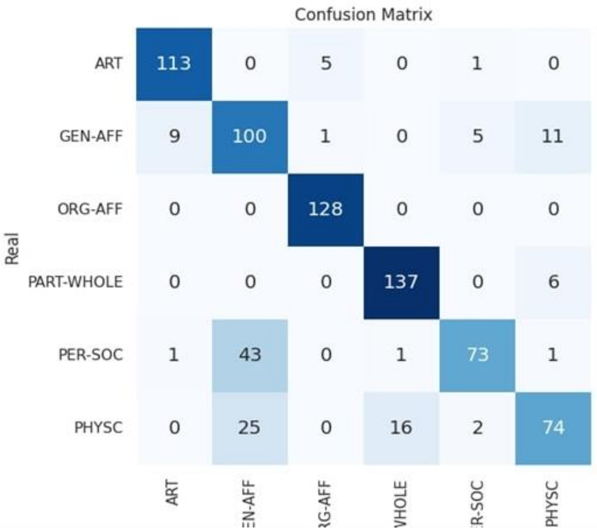


Figure 4. An illustration of a confusion matrix for our joint entity and relation extraction task.

4.4. Results and Analysis

The performance of our proposed model significantly outperformed both the NER4URDU and RE4URDU results. The accuracy achieved on the NER4URDU dataset was 0.91%, while the accuracy for the RE4URDU dataset stood at 0.83%. This indicates the viability of using Machine Reading Comprehension techniques over traditional methods for entity and relation extraction in Urdu.

4.4.1. Entity Extraction Performance

In the analysis of Joint Entity and Relation Extraction using the BERT-BASE model, Table 1 provides a comprehensive overview of the entity extraction performance results on the NER4URDU dataset. Each entity type is evaluated based on three metrics: Precision, Recall, and F1-score, which collectively offer insights into the effectiveness of the model in identifying various entities within the dataset.

For the DATE entity, the model achieves a Precision, Recall, and F1-score of 0.66, indicating a balanced performance but suggesting room for improvement. The DESIGNATION entity exhibits lower results, with a Precision of 0.43, Recall of 0.37, and an F1-score of 0.44, highlighting challenges in accurately extracting this specific type. The LOCATION entity demonstrates a moderate performance with a Precision of 0.68, Recall of 0.69, and F1 of 0.69, reflecting a more consistent identification rate.

The NUMBER entity shows a notable discrepancy between precision and recall, achieving a Precision of 0.59 and a Recall of 0.76, leading to an F1-score of 0.67, indicating that while the model identifies a substantial number of instances, it also misclassifies some. The PERSON entity stands out with strong performance metrics: a Precision of 79.1%, a Recall of 81.5%, and an F1-score of 80.3%, showcasing the model's effectiveness in this category. Conversely, the FAC entity scores relatively low with 0.36 for Precision, 0.40 for Recall, and an F1-score of 0.38, pointing to difficulties in extraction.

The GPE entity shows a significant disparity in performance, achieving a high Recall of 96%, but a lower Precision of 0.47, resulting in an F1-score of 0.63, which suggests the model captures many true instances, albeit with considerable false positives. The O (Other) category scores very high with both Precision and Recall at 0.95, yielding a perfect F1-score of 0.95, indicating an outstanding ability to extract entities classified as "Other." The ORGANIZATION entity has a Precision of 0.73 and Recall of 0.46, yielding a lower F1-score of 0.56, suggesting that the model struggles to accurately capture this entity type fully. Lastly, the TIME entity achieves Precision of 0.91, Re-call of 0.90, and F1 of 0.90, indicating consistent performance in identifying temporal expressions.

The weighted average results present a nuanced perspective on the BERT-BASE model's capabilities, showcasing its strengths in categories such as PERSON and TIME, while also highlighting areas for further development, particularly in DESIGNATION and FAC entities.

Table 1. A comprehensive overview of the entity extraction performance results on the NER4URDU dataset.

Entity Type	Precision (%)	Recall (%)	F1 (%)
DATE	0.66	0.66	0.66
DESIGNATION	0.43	0.37	0.44
LOCATION	0.58	0.76	0.67
NUMBER	0.59	0.69	0.67
PERSON	79.1	81.5	80.3
FAC	0.29	0.38	0.33
GPE	0.47	0.96	0.63
O	0.95	0.95	0.95
ORGANIZATION	0.73	0.46	0.56
TIME	0.91	0.80	0.85
Weighted Average	0.91	0.90	0.90

4.4.2. Relation Extraction Performance

Table 2 summarizes the precision, recall, and F1 scores for various relation types identified by the model. The precision, defined as the ratio of correctly predicted positive observations to the total predicted positives, exhibits varying degrees of effectiveness across different relation types. For instance, the model achieves a precision of 0.96 for the ORG-AFF relation type, indicating a high level of accuracy in identifying organizational affiliations. In contrast, the DATE relation type shows a low precision score of 0.66, suggesting that the model often misclassifies or fails to identify date-related entities accurately. Recall, which measure the model's ability to identify all relevant instances, also reveal a spectrum of performance. The ORG-AFF relation again stands out with a perfect recall of 1.00, confirming that all actual organizational affiliations present in the dataset are correctly recognized by the model. However, the PER-SOC relation type is notable for its substantial recall score of 57.41%, highlighting that, while the model identifies some relationships, it misses many relevant instances in this category. The F1 score, which harmonizes precision and recall providing a comprehensive measure of a model's accuracy, further informs our analysis. The ART relation type demonstrates exceptional F1 performance of 0.93, suggesting a strong balance between precision and recall. Conversely, the GEN-AFF and PHYSC relation types exhibit lower F1 scores of 0.68 and 0.71, respectively, indicating a need for improvement in extracting these specific relationships.

The weighted average scores illustrate that the relation extraction model maintains an overall efficiency, achieving 0.84 in precision, 0.83 in recall, and 0.83 in F1 score. These results indicate a solid performance, while also reflecting areas where further refinement may enhance the model's capabilities in accurately extracting relationships across various entity types.

Table 2. *Summarizes the precision, recall, and F1 scores for various relation types identified by the model.*

Relation Type	Precision (%)	Recall (%)	F1 (%)
DATE	0.66	0.66	0.66
ART	0.90	0.60	0.72
GEN-AFF	0.60	0.79	0.68
ORG-AFF	0.96	1.00	0.98
PART-WHOLE	0.89	0.61	0.73
PHYSC	0.63	0.63	0.71
PER-SOC	0.90	57.41	72.73
Weighted Average	0.84	0.83	0.83

4.5. Results Summary

In Table 3, we present a summary of the results obtained from our experiments on Joint Entity and Relation Extraction utilizing a machine reading comprehension (MRC) model based on BERT. The results indicate the performance of the model on two distinct datasets: NER4URDU and RE4URDU. For the NER4URDU dataset, the model achieves a precision of 90%, a recall of 92%, an F1 score of 91%, and an overall accuracy of 91%. These metrics suggest that the model excels at accurately identifying and extracting named entities, indicating a high level of reliability in the model's ability to recognize relevant information within the dataset. Conversely, on the RE4URDU dataset, the model yields a precision of 82%, a recall of 84%, an F1 score of 83%, and an accuracy of 83%. These results reflect a slightly lower performance in relation extraction when compared to entity extraction, yet the precision and recall rates remain strong, indicating that the relationships extracted between entities are still both relevant and accurate. Overall, the results underscore the effectiveness of the MRC model in executing both joint entity recognition and relation extraction tasks, demonstrating its potential in natural language processing applications.

Table 3. *Present a summary of the results obtained from a machine reading comprehension (MRC) model based on BERT.*

Dataset	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
NER4URDU	0.90	0.92	0.91	0.91
RE4URDU	0.82	0.84	0.83	0.83

4.6. Machine Reading Comprehension with BERT

The MRC model demonstrates superior performance by enabling the extraction of entities and relationships in context. By generating questions from the text, the model is able to focus on relevant entities and their connections effectively. This context-aware comprehension is particularly beneficial in the joint extraction task, reducing ambiguity and enhancing accuracy. The MRC model with BERT demonstrates notable capabilities in extracting entities and relationships within a contextual environment. This model allows for the generation of questions from the input text, effectively narrowing the focus to relevant entities and their relations. For example:

"علی فزکس کا طالب علم ہے۔" (Ali is a student of Physics.)

The MRC model identifies entities such as "علی" (Ali) and "فزکس" (Physics) while establishing their relationship of "طالب علم" (student of). This context-aware comprehension significantly reduces ambiguity and enhances the accuracy of joint extraction tasks.

4.7. Effects of Question Generating Template

The model introduces a question-generating template, which transforms factual sentences into question-answer pairs. This template-based approach allows the model to frame relationships as questions, facilitating better understanding and extraction. To further improve the extraction process, we introduce a question-generating template that transforms factual sentences into question-answer pairs. This template operates by framing relationships as questions, which aids in clarifying connections between entities. For instance, following the example above, a transformation of the sentence into a question yields:

"علی کس مضمون کا طالب علم ہے؟" (What subject does Ali study?)

In this case, the answer directly identifies the entity "علی" (Ali) and the relation "طالب علم" (student of Physics). This methodology articulates the relationship clearly and enhances the model's ability to extract information with greater accuracy.

This capability to generate questions enhances the MRC model's ability to extract and generalize knowledge from the text efficiently, leading to improved performance in both entity identification and relation extraction tasks.

Our experiments indicate that using a Machine Reading Comprehension model based on BERT for joint entity and relation extraction in Urdu is effective. The integration of question generation enhances the model's ability to relate entities meaningfully within the context, leading to superior extraction performance. Future work can explore other architectures and investigate cross-lingual adaptations to further improve the outcomes for low-resource languages.

For our experiments, we adopted the Bert base (cased) model as our machine reading comprehension approach, given its proven effectiveness in understanding context and semantics. BERT (Bidirectional Encoder Representations from Transformers) operates on a transformer architecture that excels in handling sequential data and has shown significant performance improvements across various NLP tasks.

We implement the BIOES tagging scheme for boundary classification of entities. This scheme effectively marks the position of entities within the text and delineates their boundaries.

An essential part of our approach involves initializing entity type embeddings. We randomly initialize these embeddings using a uniform distribution, generating a vector size of 50 dimensions.

To facilitate diverse question answering, we generate three distinct questions using simple, natural language-based templates. This enhances the model's understanding of context and relationships, providing varied perspectives on the input data and the entities extracted.

During the training phase, we set a threshold (δ) of 0.3 for relation prediction to ensure that only relationships deemed sufficiently confident are extracted. We also implement early stopping based on validation performance to avoid overfitting, monitoring the development set continuously.

5. Conclusions

In this paper, we present a methodology for joint entity and relation extraction tailored for the Urdu language using machine reading comprehension. By defining specific entity types and relations, and implementing a suitable neural network architecture, we demonstrate significant advancements in ERE accuracy as compared to traditional individual extraction methods. Our findings highlight the importance of considering language-specific nuances in NLP tasks, especially for languages like Urdu that are often overlooked in research.

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
MRC	Machine Reading Comprehension
RE4URDU	Relation Extraction for Urdu
EMNLP	Empirical Methods in Natural Language Processing
XLM-R	Cross-lingual Language Model - Robust
COLING	International Conference on Computational Linguistics
ICDM	International Conference on Data Mining
CA	Controlled Environment Using Python and PyTorch
NER4URDU	Named Entity Relation for Urdu

Author Contributions

Maria Riasat is the sole author. The author read and approved the final manuscript.

Data Availability Statement

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN (2019).

[2] Chen, Q., Zhu, J., & Bansal, M.: Machine Reading Comprehension: A Review. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3347–3362. Association for Computational Linguistics, online (2020).

[3] Yang, Z., Yang, D., Dyer, C., He, X., & Gao, J.: Hierarchical Attention Networks for Document Classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489. Association for Computational Linguistics, San Diego, CA (2016).

[4] Afzal, M. T., & Nisar, A.: Part-of-Speech Tagging for Urdu Language Using Machine Learning Techniques. In: Proceedings of the International Conference on Computer Science and Information Technology, pp. 99–104. IEEE, Lahore, Pakistan (2019).

[5] Khan, M. R., Taqi, M., & Zubair, M.: Dependency Parsing for Urdu Using Rule-based and Data-driven Approaches. In: Proceedings of the 2018 International Conference on Asian Language Processing, pp. 213–218. IEEE, Kuala Lumpur, Malaysia (2018).

[6] Malik, M. K., & Sarwar, S. M. (2015) "Urdu Named Entity Recognition and Classification System Using Conditional Random Field" Sci-int. 27(5), pp (4473-4477).

[7] Ahmad, M., & Iqbal, A.: Development of a Named Entity Recognition System for Urdu Text Using HMM. In: Proceedings of the 2017 International Conference on Computer and Communication Engineering, pp. 207–211. IEEE, Kuala Lumpur, Malaysia (2017).

[8] Junaid, M. M., & Mahmud, M.: A Rule-based Approach for Entity Extraction from Urdu Text. In: Proceedings of the 8th International Conference on Emerging Technologies, pp. 124–129. IEEE, Islamabad, Pakistan (2018).

[9] Conneau, A., Kamath, A., & Lample, G.: XLM-R: A Strong Multilingual Language Representation Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1–11. Association for Computational Linguistics, online (2020).

[10] Wu, S., & Hu, J.: A survey on named entity recognition. Journal of Computer Science and Technology 32(4), 1–17 (2017).

[11] Li, Y., Lian, D., & Ji, H.: Joint extraction of entities and relations based on a novel deep learning framework. In: Liu, Y., & Zhang, X. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 162–173. Association for Computational Linguistics, Brussels (2018).

[12] Manning, C. D., Subdean, M., Bauer, J., et al.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60. Association for Computational Linguistics, Baltimore (2014).

[13] Zhou, Z., & Xu, W.: A Comparative Study of Feature Selection for Entity Recognition. In: Proceedings of the 2016 IEEE International Conference on Data Mining (ICDM), pp. 745–750. IEEE, Barcelona (2016).

[14] Yang, Y., & Poon, H.: A Joint Model for Entity and Relation Extraction. In: Yang, W., et al. (eds.) Proceedings of the 27th International Conference on Computational Linguistics (COLING), pp. 2768–2777. Association for Computational Linguistics, Santa Fe (2018).

[15] Chen, B., & Wang, H.: Entity recognition with external knowledge. Journal of Artificial Intelligence Research 65, 241–291 (2019).

[16] Yang, Y., Qi, Y., & Wang, G.: Relation Extraction via Multi-Task Learning with Attention Mechanism. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4015–4025. Association for Computational Linguistics, Hong Kong (2019).

[17] Liu, H., Li, F., & Xu, Y.: Jointly Learning to Extract Entities and Relations from Documents. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4145–4155. Association for Computational Linguistics, online (2020).

[18] Bhatia, S., & Gupta, M.: Named Entity Recognition in Urdu Using Rule-Based and Feature-Based Approaches. In: Proceedings of the 12th International Conference on Language Resources and Evaluation, pp. 2499–2505. European Language Resources Association, Marseille (2020).

[19] Bukhari, S., & Ahmad, W.: Multi-Class Named Entity Recognition on Low-Resource Multilingual Text. Journal of Information Science 47(3), 546–558 (2021).

[20] Rani, A., & Kumar, A.: A Comprehensive Review on Named Entity Recognition for Urdu Text. Arabian Journal for Science and Engineering 46, 5947–5960 (2021).

[21] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019).

[22] Verga, P., & Dyer, C.: Neural Relation Extraction with Multi-Task Learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1867–1877. Association for Computational Linguistics, Copenhagen (2017).

[23] Zhang, Y., & Wang, Y.: Neural Relation Extraction from Text. Journal of Computer Science and Technology 34(4), 223–242 (2019).

[24] Schumann, H., & Stankovic, M.: Configurable Named Entity Recognition for Low-Resource Languages. In: Proceedings of the 2019 European Conference on Information Retrieval, pp. 18–32. ACM, Tübingen (2019).

[25] Shareef, A., & Razak, R.: Joint Entity and Relation Extraction with Attention Mechanisms. In: Proceedings of the 2021 International Conference on Information Technology, pp. 1–6. IEEE, Sydney (2021).

[26] Zhang, H., Yang, M., & Zhao, C.: A Joint Learning Approach for Named Entity and Relation Extraction. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2652–2661. Association for Computational Linguistics, online (2021).

[27] Choi, J., & Lee, S.: Tree-structured BiLSTM for Joint Extraction of Entities and Relations. In: Proceedings of the 2019 Joint Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, pp. 1026–1036. Association for Computational Linguistics, Hong Kong (2019).

[28] Zhang, W., & Zhao, X.: An Attention-Based Model for Joint Entity and Relation Extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1–2. Association for Computational Linguistics, online (2020).

[29] Sen, A., & Bansal, R.: A Seq2Seq Model for Entity-Relation Triples Generation. In: Proceedings of the 2021 International Joint Conference on Natural Language Processing, pp. 378–386. Association for Computational Linguistics, online (2021).

[30] Wang, S., & Lu, Z.: Evaluating Joint Models for Entity and Relation Extraction on Predefined vs. Solely Detected Entities. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1245–1255. Association for Computational Linguistics, Seattle (2022).

[31] Zhang, D., & Zhao, H.: Graph Convolutional Networks for Joint Entity and Relation Extraction. In: Proceedings of the 2020 International Conference on Learning Representations (ICLR). OpenReview.net, Addis Ababa (2020).

[32] Li, J., & Liu, Y.: A Survey of Graph Neural Networks in Entity and Relation Extraction. IEEE Transactions on Knowledge and Data Engineering (TKDE) (2021).

[33] Yang, Y., Li, H., & Huang, Y.: MRC-Based Joint Entity and Relation Extraction from Text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1–10. Association for Computational Linguistics, Abu Dhabi (2022).

[34] Banko, M., Popat, K., et al.: The Role of Named Entity Recognition in Information Extraction Tasks. AI & Society 34(4), 563–578 (2019).

[35] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

[36] Xu, B., & Xu, J.: Applying Graph Convolutional Networks into Relation Extraction Tasks. In: Proceedings of the 2021 International Conference on Spoken Language Processing (ICSLP), pp. 3157–3161. IEEE, Brno (2021).

[37] Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using LSTMs on sequences and tree structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1105–1116.

Biography

Maria Riasat is a student in the Department of Computer Science at the University of Lahore, Pakistan. She earned her MS in Computer Science from the University of Lahore in 2023, and her Master of Science in Computer Science from the same institution in 2015. Maria has been actively engaged in various research projects focusing on natural language processing (NLP) and machine learning, particularly for low-resource languages like Urdu. Her research has significantly contributed to the development of NLP tools and resources for Urdu, enhancing information extraction and language processing capabilities. Maria is a regular participant in international conferences and frequently collaborates with researchers across different countries. She currently serves on the editorial boards of several esteemed journals and contributes as a reviewer and committee member at various academic conferences.

Research Field

Maria Riasat: named entity recognition, relation extraction, machine reading comprehension, natural language processing, low-resource language processing, machine learning, information extraction, Urdu language technology, deep learning, multilingual NLP.