Event-Arguments Extraction Corpus and Modeling using BERT for Arabic

Alaa Aljabari
Birzeit University
aaljabari@birzeit.edu

Lina Duaibes
Birzeit University
1205358@student.birzeit.edu

Mustafa Jarrar
Birzeit University
mjarrar@birzeit.edu

Mohammed Khalilia
Birzeit University
mkhalilia@birzeit.edu

Abstract

Event-argument extraction is a challenging task, particularly in Arabic due to sparse linguistic resources. To fill this gap, we introduce the ***Wojood***^{Hadath} corpus (550k tokens) as an extension of Wojood, enriched with event-argument annotations. We used three types of event arguments: *agent*, *location*, and *date*, which we annotated as relation types. Our inter-annotator agreement evaluation resulted in 82.23% *Kappa* score and 87.2% F_1-score. Additionally, we propose a novel method for event relation extraction using BERT, in which we treat the task as text entailment. This method achieves an F_1-score of 94.01%. To further evaluate the generalization of our proposed method, we collected and annotated another out-of-domain corpus (about 80k tokens) called ***Wojood***$^{\text{OutOfDomain}}$ and used it as a second test set, on which our approach achieved promising results (83.59% F_1-score). Last but not least, we propose an end-to-end system for event-arguments extraction. This system is implemented as part of SinaTools, and both corpora are publicly available at https://sina.birzeit.edu/wojood

1 Introduction

Understanding and extracting events from text is crucial in natural language understanding (Khalilia et al., 2024) for applications like disaster monitoring (Hernandez-Suarez et al., 2019), emergency response (Simon et al., 2015), insurance decision support, and fostering community resilience (Ahmad et al., 2019). Events, a type of named entity mentions (Jarrar et al., 2023a), are connected to other entities through their arguments. This connection forms the foundation of the event-argument extraction task, which is closely related to relation extraction. By identifying events and linking them with arguments like agents, locations, and dates, we establish meaningful relationships that enhance applications such as information retrieval systems (Singh, 2018), word sense disambiguation (Jarrar et al., 2023b; Al-Hajj and Jarrar, 2021), and knowledge graph construction (Ye et al., 2022). Figure 1 shows how the event "The Economic Forum Meeting" is connected with its three arguments.

Figure 1: An event annotated with its arguments.

Despite the significance of event-argument extraction, there is a notable gap in the availability of comprehensive, annotated corpora for this purpose, especially for under-resourced languages like Arabic (Darwish et al., 2021; Haff et al., 2022). To address this gap, we have developed an event-argument extraction corpus specifically for Arabic. We extended the Wojood corpus (Jarrar et al., 2022a), which is the largest and most recent NER corpus for Arabic (Liqreina et al., 2023; Jarrar et al., 2024, 2023a). We annotated arguments of event entities in Wojood corpus by manually linking each event entity with its corresponding argument entities. As shown in Figure 2, three types of arguments are introduced (*agent*, *location*, *date*) for each event. Our Wojood extension (i.e., the event-arguments annotations) is called ***Wojood***^{Hadath}.

Furthermore, we introduce a novel method for event-argument extraction using BERT and achieved 94.01% F_1-score. Based on ***Wojood***^{Hadath}, we generated a dataset of premise-hypotheses sentence pairs (we call it ***Hadath***$^{\text{NLI}}$). We used this dataset to fine-tune BERT, as a natural language inference (NLI) task. To test the generalization of our approach, we have constructed an additional out-of-domain dataset (about 80k tokens) called ***Wojood***$^{\text{OutOfDomain}}$ and used it as a second test set. Our model achieves,

again, promising results (83.59% F_1-score). Finally, to streamline the event-argument extraction, we propose an end-to-end system specifically designed for the Arabic event-argument extraction task. In summary, the contributions of this paper are:

1. ***Wojood***^{Hadath} corpus (550k tokens) manually annotated with event argument relations. This corpus is used to generate an NLI dataset ***Hadath***$^{\text{NLI}}$ (25k premise-hypotheses pairs).

2. ***Wojood***$^{\text{OutOfDomain}}$, an additional annotated corpus (80k tokens) for out-of-domain evaluation.

3. Novel methodology for event-argument extraction by framing the task as an NLI problem, achieving high performance.

4. Novel end-to-end system for event-argument relation extraction.

All datasets are available online[1], and the end-to-end system is implemented and can be used as part of the SinaTools (Hammouda et al., 2024)[2].

This paper is organized as follows: Section 2 reviews previous research. Section 3 discusses corpus annotations. Section 4 presents the inter-annotator agreement analysis. Section 5 outlines the methodology. Section 6 details the dataset construction, while Section 7 covers the experimental setup. Section 8 introduces the out-of-domain dataset. Section 9 elaborates on the end-to-end system, and Section 10 explores ablation studies. Finally, Section 11 concludes the paper and discusses future directions.

2 Related works

Events are occurrences or actions that happen over time, involving specific participants and locations. They have temporal components and rely on physical entities to take place (Jarrar, 2021; Jarrar and Ceusters, 2017). Extensive research has focused on event extraction and event argument extraction (EAE), especially in English. However, EAE in Arabic is limited, leaving a gap in the literature (Chouigui et al., 2018; Alomari et al., 2020).

Automated methods, utilizing either statistical algorithms or NLP techniques, are employed to identify relationships between words, using large corpora to detect related terms through their co-occurrence patterns (Khallaf et al., 2023). For instance, Hkiri et al. (2016) proposed a model to extract event entities from Arabic news articles using the GATE tool, employing a five-stage entity identification process that establishes links between events and their corresponding arguments. However, their dataset is small as it consists of only 1,650 sentences.

In a similar context, AL-Smadi and Qawasmeh (2016) proposed to use an unsupervised rule-based technique to extract events and the relationship with their associated entities from 1,000 Arabic tweets covering Time, Agent, Location, Target, Trigger, and Product. They linked extracted events and entities to a knowledge base, achieving 75.9% accuracy. This accuracy pertains to the textual representation within the tweet corresponding to the event expression, event type identification (97.7% accuracy), and event time extraction (87.5% accuracy). However, despite these achievements, there remains a need for larger and more diverse datasets to further validate these findings.

Multilinguality has increasingly attracted attention across various fields due to its potential to enhance the understanding and processing of diverse linguistic data (Jarrar and Amayreh, 2019; Jørgensen et al., 2023; Duaibes et al., 2024). This direction has notably influenced relation extraction tasks, as seen with SMILER (Seganti et al., 2021), which aims to improve entity and relation extraction, including Arabic. Seganti et al. (2021) fine-tuned HERBERTa on the Arabic subset of SMILER, achieving high performance in identifying relations and entities despite its smaller dataset of 9k sentences. This suggests that the unique linguistic features of Arabic may contribute to the model's robustness. Furthermore, Cabot et al. (2023) introduced valuable resources, namely, SRED$^{\text{FM}}$ and RED$^{\text{FM}}$, designed for multilingual relation extraction. SRED$^{\text{FM}}$ encompasses over 40 million triplet instances across 18 languages, featuring 400 relation types and 13 entity types. Their mREBEL model, pre-trained on SRED$^{\text{FM}}$, exhibited a remarkable improvement of 15 points in Micro-F_1 compared to HERBERTa. However, their qualitative error analysis excludes Arabic and Chinese because the authors are not proficient in these languages.

Joint extraction models have become increasingly important in NLP. Addressing the need for

[1]Datasets: https://sina.birzeit.edu/wojood/

[2]SinaTools: https://sina.birzeit.edu/sinatools

improved event extraction in Arabic text, El Khbir et al. (2022) introduced a joint model for event extraction in Arabic text using the ACE 2005 dataset. They used a graph-based representation to extract entities, relationships, and event triggers, along with their arguments. This approach led to improved Arabic NER accuracy through different experiments and tokenization schemes.

In the realm of biomedicine, Xu et al. (2022) introduced the NBR (NLI improved Biomedical Relation Extraction) method, which verbalizes relations in natural language hypotheses, allowing the model to utilize semantic information effectively in prediction, even with limited data. In contrast, Cao et al. (2023) addresses cross-lingual EAE challenges through their innovative LanguAge-oriented Prefix-tunIng Network (LAPIN) approach. LAPIN utilizes a language-oriented prefix generator module to handle language variations and a language-agnostic template constructor module to create adaptable templates. Experimental results demonstrate LAPIN's outperforming, achieving an average F_1 improvement of 4.8% and 2.3% on two multilingual EAE datasets compared to the previous state-of-the-art models. These efforts collectively contribute to advancing the field of EAE in various languages, including Arabic.

3 Dataset and annotation

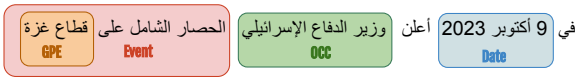
3.1 Corpus Preparation

To construct an event-argument corpus, we extended the existing Wojood corpus (Jarrar et al., 2022b). This is because it is a rich Arabic nested named entity corpus comprising 550k tokens, supporting 21 distinct entity types, including categories such as person, organization, location, event, and date. Notably, it incorporates 2,772 annotated events. According to Wojood guidelines, a mention is annotated as an event if it represents an occurrence of general interest, like battles, wars, sports events, demonstrations, disasters, elections, and national or religious holidays.

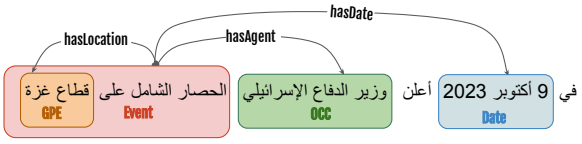
Our objective in this paper is to identify event arguments and establish relationships between these arguments and the respective event entities, as illustrated in Figure 2.

3.2 Annotation Process

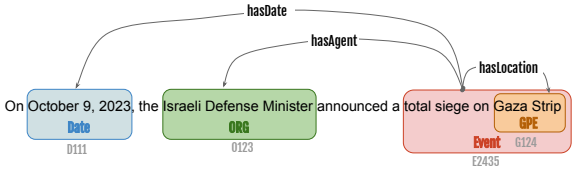
Initially, we assigned a unique identifier to each entity in Wojood, see Figure 2(a). Then, we linked these entities with relationships, see Figure 2(b).



(a) Annotated entities in Wojood.



(b) Event with argument entities annotated as relations.



(c) Translation of the above example.

Figure 2: Annotating an event entity with its arguments.

3.2.1 Relationship Types

We propose to use these relations:

- $hasAgent$: specifies participant(s) involved in the event, which can be a PERS, ORG, OCC, or NORP named entities.
- $hasLocation$: indicates where the event occurred, which can be GPE, LOC, and FAC named entities.
- $hasDate$: points when the event occurred, which can be TIME or DATE.

3.3 Annotation Guidelines

We propose the following guidelines to annotate the corpus:

1. Event arguments are recognized only within the same sentence.
2. Entities with different entity IDs are considered distinct entities. For example, in the sentence (مقتل الرئيس المصري أنور السادات / The killing of the Egyptian president Anwar al-Sadat), the entity (الرئيس المصري/the Egyptian president) and the entity (أنور السادات / Anwar al-Sadat) are regarded as separate entities, thus two agents in the event. However, in reality, they refer to the same individual.

3. The same event may have multiple agents, as in (توقيع اتفاقية تعاون بين الحكومة اللبنانية و البنك المركزي / Signing a cooperation agreement between the Lebanese government and the Central Bank). In Figure 3, (الحكومة اللبنانية / the Lebanese government) and (البنك المركزي / the Central Bank) are both agents of the same event.

4. In the same sentence, two different event entities can share the same argument. For example, in the sentence (الوضع السياسي متوتر في مصر بعد حرب عام ١٩٦٧ (النكسة)/The political situation in Egypt is tense after the 1967 War (Al Naksa)), there are two event entities, according to Wojood guidelines. The first event is (حرب عام ١٩٦٧/The 1967 war), and the second event is (النكسة/Al Naksa). In this way, the entity (عام ١٩٦٧/the year 1967) is considered the argument for both events in the sentence.

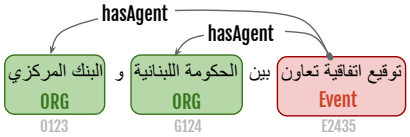


Figure 3: An event with more than one agent

3.4 Corpus Statistics

Wojood comprises 550k tokens with 2,772 event entities. Among these entities, we annotated 1,974 events with event-arguments relations - the other 798 do not have arguments. Notably, there are 355 (18%) events that are annotated with at least two arguments, and 77 (18% of agent relations) annotated with multiple agents. This underscores the corpus's rich and diverse interconnection of events and their roles. The number of instances for each relation type is shown in Table 1.

Relation	Count
hasAgent	423
hasLocation	833
hasDate	1,332
Total	**2588**

Table 1: Number of relations in ***Wojood***^{Hadath}

4 Inter-annotator Agreement

To evaluate the quality of our annotations, we randomly selected 5% of the annotations and asked our two annotators to annotate in parallel. We computed the Inter-annotator Agreement (IAA) using Cohen's *kappa* and F_1-score. The results in Table 2 illustrate high agreement.

4.1 Calculating *kappa*

We assessed annotator agreement for three relations. Agreement happens when both annotators assign the same relationship type. The *kappa* coefficient (Eugenio and Glass, 2004) is defined as:

$$K = \frac{P_0 - P_e}{1 - P_e} \quad (1)$$

where P_o is the observed agreement, and P_e is the expected agreement. P_e is calculated as :

$$P_e = \frac{1}{N^2} \sum_T n_{T1} \times n_{T2} \quad (2)$$

4.2 Calculating F_1-score

The F_1-score for a specific relation (e.g., $hasAgent$) is calculated using Equation 3. True positives (TP) are those when annotators agree, while false negatives (FN) and false positives (FP) arise from disagreements. FN arises from the first annotator's disagreement, while FP arises from the second annotator's disagreement.

$$F_1 - Score = \frac{2TP}{2TP + FN + FP} \quad (3)$$

4.3 Discussion and Annotation Challenges

We encountered several challenges, including:

1. Although event arguments should be encompassed within the sentence referencing the event, it was not always easy for the annotators to decide if entities within the same sentence as the event should be annotated as event entities.

2. In the $hasAgent$ relation, it can sometimes be challenging to determine whether an entity serves as an agent for the event. See the example in Figure 4.

Figure 4: Example of disagreement: whether the "Secretary-General" is an agent in the sentence (*letter by the Secretary-General on the occasion of World Environment Day*).

Relation	TP	FN	FP	κ	F1-Score
$hasAgent$	37	10	10	67.85%	79%
$hasLocation$	29	2	2	91.70%	94.00%
$hasDate$	43	2	6	87.15%	91%
Overall	**109** count	**14** count	**18** count	**82.23%** macro	**87.20%** micro

Table 2: Overall IAA for each relation.

3. The annotators have different levels of experience; one is skilled at defining event arguments, while the other is not. Consequently, one annotator agreed to designate a specific entity as the agent, while the second did not.

5 Even-Argument Extraction (EAE)

In this section, we provide an in-depth explanation of our approach to addressing the EAE task through the NLI task. Figure 5 illustrates our framing of the EAE problem.

5.1 Problem Formulation

The objective of the EAE task is to identify the relationships between the event(s) in a sentence and the named entities mentioned in the same sentence. That is, given a sentence s annotated with a set of event entities $E = \{e_i\}_{i=1}^{n}$ and a set of other named entities $N = \{n_j\}_{j=1}^{m}$, the goal of EAE is to identify the relation $r \in R$ for each pair (e_i, n_j) in s, where $R \in \{hasAgent, hasLocation, hasDate\}$.

5.2 Event Relation Extraction as NLI

We propose to solve the EAE by framing it as a Natural Language Inference (NLI) task. In NLI, we assess whether one sentence (the *premise*) entails another (the *hypothesis*). That is, a pair of sentences is classified as $True$ or $False$. To extract an event argument relation from a sentence (See Figure 5), we treat the original sentence as a premise and generate the hypothesis automatically. The hypothesis is another sentence generated using a template, to represent a possible relationship. In other words, we propose to treat EAE as a binary NLI task focusing on entailment. The input sentence s is the premise, while the hypothesis is a verbalized template representing a relation r between event e_i and a named entity n_j mentioned in the same sentence. The model then determines if the premise "entails" the hypothesis (as classification $True/False$), indicating the existence of the relation r between e_i and n_j.

Template Construction

For each of the three relations, we designed a template parameterized over the named entities mentioned within the same sentence. Each template has two placeholders: Event Placeholder P_{event} and Entity Placeholder P_{entity}. These placeholders are filled with event mention e_i and named entity mention n_j, respectively. Figure 5 illustrates how a template can be used to generate a hypothesis for the $hasDate$ relation between the (الحصار الشامل على قطاع غزة / Total siege on Gaza strip) event and the (٩ أكتوبر ٢٠٢٣ /October 9, 2023) entity. The three templates used are shown in Table 3

Sentence Encoder

The sentence encoder is used to extract representations for input text. In our approach, the input text is a pair consisting of a premise (the input sentence s) and a hypothesis (the filled-in template h). A transformer encoder, denoted by T, is used to process the input sequence and derive its representation, $T \rightarrow \mathbb{R}^d$. This encoder effectively captures contextual information and semantic relationships within the text pair generating representation H for the sentence pairs. Formally,

$$H = T\left([CLS]\mathbf{s}[SEP]\mathbf{h}\right) \tag{4}$$

where $[CLS]$ is the special classification token, $[SEP]$ is a separator between the premise and hypothesis, and d is the encoding dimension.

Relation Classifier

The feature vector H from the sentence encoder is input into a fully connected layer. This layer calculates the predicted probability that the premise entails the hypothesis, with a $True$ result indicating that the relationship r specified in h exists between the event e and the entity n. Formally,

$$\hat{y}_i = \sigma\left(HW + b\right) \tag{5}$$

where $\hat{y}$ indicates whether the hypothesis holds a positive relation, while W and b represent a weight matrix and a bias term, respectively.

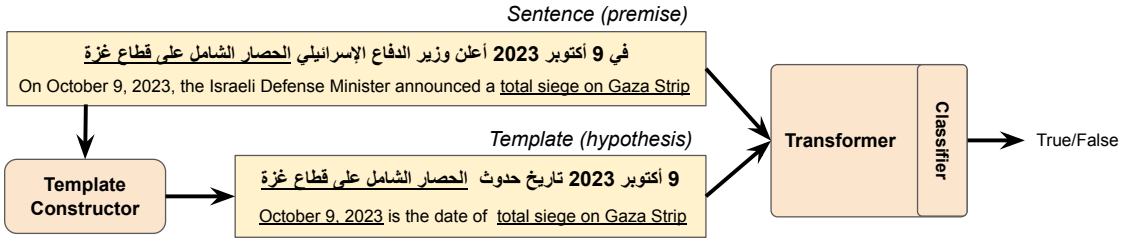


Figure 5: Framing the EAE task as NLI task.

Relationship	Template	Example
$hasAgent$	P_{event} أحد الفاعلين في P_{entity}	وزير الدفاع الإسرائيلي أحد الفاعلين في الحصار الشامل على قطاع غزة The Israeli defense minister is an agent in Total siege on Gaza strip
$hasLocation$	P_{event} مكان حدوث P_{entity}	قطاع غزة مكان حدوث الحصار الشامل على قطاع غزة Gaza strip is place of occurring Total siege on Gaza strip
$hasDate$	P_{event} تاريخ حدوث P_{entity}	٩ أكتوبر ٢٠٢٣ تاريخ حدوث الحصار الشامل على قطاع غزة October 9,2023 is the date of occurring Total siege on Gaza strip

Table 3: Templates utilized during the testing phase.

Training Objective

Our training objective is to prioritize the accurate identification of positive instances over negatives. To accomplish this, we utilize a weighted cross-entropy loss L_{WCE} (Eq. 6), which penalizes misclassifications of positive instances more heavily. Additionally, we use Noise Contrastive Estimation Loss L_{NCE} (Eq. 7) to improve the discrimination between positive and negative instances (Robinson et al., 2020). The final loss function is expressed as $Loss = L_{WCE} + L_{NCE}$.

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\Big(w_p y_i \log(\hat{y}_i) + w_n(1-y_i)\log(1-\hat{y}_i)\Big) \quad (6)$$

$$L_{NCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(s(y_i)/\tau)}{\sum_j \exp(s(y_j)/\tau)}\right) \quad (7)$$

6 Construct *Hadath*$^{\text{NLI}}$ Dataset

Based on our annotated ***Wojood***^{Hadath} corpus, we created an NLI dataset (called ***Hadath***$^{\text{NLI}}$), which is a dataset of premise-hypotheses sentence pairs that we use to train the EAE model. To create this dataset, we carried out the following steps:

1. Premise Sentences Preparation

We extracted sentences from the ***Wojood***^{Hadath} corpus and utilized them as premises. These sentences were then split into training (70%) and testing (30%) sets. This split is important to avoid overlap between premises in the two splits during the generation of positive and negative pairs.

2. Hypothesis Sentences Preparation

Multiple hypotheses are generated for each premise sentence, utilizing the entities mentioned within the sentence. These entities are annotated as arguments of a specific event in the ***Wojood***^{Hadath} corpus. That is, based on the event-argument relation annotations (see section 3.2.1), each event and its arguments are used to generate a hypothesis using the template for relation r. This hypothesis is then paired with the premise sentence to form a **positive** pair.

3. Generating Negative Pairs

In premise sentences, entities that are not linked with events (i.e., not event arguments) are used to generate **negative** pairs. That is, we link events with entities that are not their correct arguments (i.e., not annotated as a relation in ***Wojood***^{Hadath}), and generate them as hypotheses using templates. Each generated hypothesis is paired with its premise and used as a negative pair.

Hadath$^{\text{NLI}}$ Dataset Statistics

For the training phase, four templates are designed for each relation, each with a particular verbalizer (see Table 5). This step aims at data augmentation and enhancing the contextual diversity of the training set. For the test phase, one template per relation was selected from the training templates to generate the hypothesis (see Table 3). The final dataset consists of $25,473$ pairs, comprising $10,478$ positive and $14,995$ negative pairs. Table 4 presents more statistics.

Phase	Pairs	Positive	Negative	Total
Train	$hasAgent$	1,248	6,156	**7,404**
	$hasLocation$	2,268	4,456	**6,724**
	$hasDate$	3,716	2,948	**6,664**
	SubTotal	**7,232**	**13,560**	**20,792**
Test	$hasAgent$	111	653	**764**
	$hasLocation$	267	464	**728**
	$hasDate$	403	318	**721**
	SubTotal	**778**	**1,435**	**2,213**
	Total	**8,010**	**14,995**	**23,005**

Table 4: Number of pairs in the ***Hadath***$^{\text{NLI}}$ Dataset

7 EAE Modeling Experiments

The ***Hadath***$^{\text{NLI}}$ is used to train an EAE model.

7.1 Training Hyperparameters

During training, we employed a k-fold strategy with $k = 5$ to ensure the robustness of model evaluation. For all experiments, we utilized the `UBC-NLP/ARBERTv2` (Abdul-Mageed et al., 2021) as the text encoder model, with a learning rate of $2e^{-5}$ and the $AdamW$ optimizer with a weight decay of $1e^{-8}$. In the L_{WCE} loss function, $w_{pos} = 1$ and $w_{neg} = .5$, while $\tau = 1$ in L_{NCE} loss function.

7.2 Results and Discussion

Table 6 shows the obtained results using the aforementioned experimental setups. We average the scores across five folds and report the model with the best average $F1score$. Table 7 provides insight into the NLI model performance in each relation. Note that this NLI performance represents the accuracy of our model in classifying sentence pairs, which is not the EAE accuracy. The accuracy of event-argument relation extraction (EAE) shall be presented in section 9.

Notably, our NLI model achieves high $F1scores$ across all event relations, with the $hasDate$ relation exhibiting the highest score (96.44%). Additionally, the negative relations also achieved a notably high F_1 score of 95.29%, indicating the model's effectiveness in recognizing entities unrelated to the event.

To validate this remarkable performance of the model and to test its generalization, we constructed a new corpus and conducted additional out-of-domain experiments.

8 Additional ***Wojood***$^{\text{OutOfDomain}}$ Dataset

To evaluate the model's generalization and its robustness to contexts beyond the ***Wojood***^{Hadath} domains, we constructed a new out-of-domain corpus (referred to as the ***Wojood***$^{\text{OutOfDomain}}$ corpus). We then extracted an NLI dataset and used it for model testing.

8.1 Corpus Preparation

This corpus covers 10 distinct domains: economics, finance, politics, science, technology, art, law, agriculture, history, and sports each containing nearly 8k tokens. The corpus covers events from 2010 to 2022, manually collected from news websites such as Aljazeera, and Alarabiya, totaling $80k$ tokens.

We used the same Wojood annotation guidelines to maintain consistency. Table 8 shows the number of instances for each relationship type.

8.2 Construct ***Wojood***$^{\text{OutOfDomain}}$ Dataset

An NLI dataset was generated from ***Wojood***$^{\text{OutOfDomain}}$ using the same methodology as ***Hadath***$^{\text{NLI}}$, employing only the templates designated for testing, resulting in a total of 1124 pairs. Detailed statistics are provided in Table 9.

8.3 Experiments and Results

The ***Wojood***$^{\text{OutOfDomain}}$ is used to evaluate the EAE model, which was trained on the ***Hadath***$^{\text{NLI}}$. The results, presented in Table 10, highlight the model's robust generalization across diverse domains, despite encountering challenges like domain-specific vocabulary and linguistic nuances. Despite a slight performance decline compared to the ***Hadath***$^{\text{NLI}}$ test set, the model maintained a high overall average $F_1-score$ of 83.38% on the ***Wojood***$^{\text{OutOfDomain}}$.

9 End-to-End System for EAE

9.1 System Architecture

This section introduces our novel end-to-end EAE system, which efficiently extracts event-related information from text by seamlessly identifying entity boundaries, determining their types, and recognizing argument entities and their relations to an event entity. We utilized the EAE NLI model to construct the system. As illustrated in Figure 6, the overall steps are:

Named Entity Recognition Module: The system starts by extracting entities and their types from the input sentence. The process is carried out through the online Wojood web service[3]. Then,

[3] https://sina.birzeit.edu/wojood

Template	*hasLocation*	*hasAgent*	*hasDate*
t_1	P_{event} (site of) هو موقع P_{entity}	P_{event} (actor of) أحد المتأثرين في P_{entity}	P_{event} (time of) هو زمن P_{entity}
t_2	P_{event} (place of occurring) مكان حدوث P_{entity}	P_{event} (agent in) أحد الفاعلين P_{entity}	P_{event} (date of occurring) تاريخ حدوث P_{entity}
t_3	P_{entity} (in) في P_{event}	P_{entity} (related with) له علاقة مع P_{event}	P_{entity} (happened at) حدث في P_{event}
t_4	P_{entity} (in) في P_{event} (happened) وقع	P_{event} (in) في P_{entity} (participate) شارك	P_{entity} (at date) بتاريخ P_{event} (happened) حدث

Table 5: Templates utilized in the training phase. Template Set t_2 is selected for the testing phase.

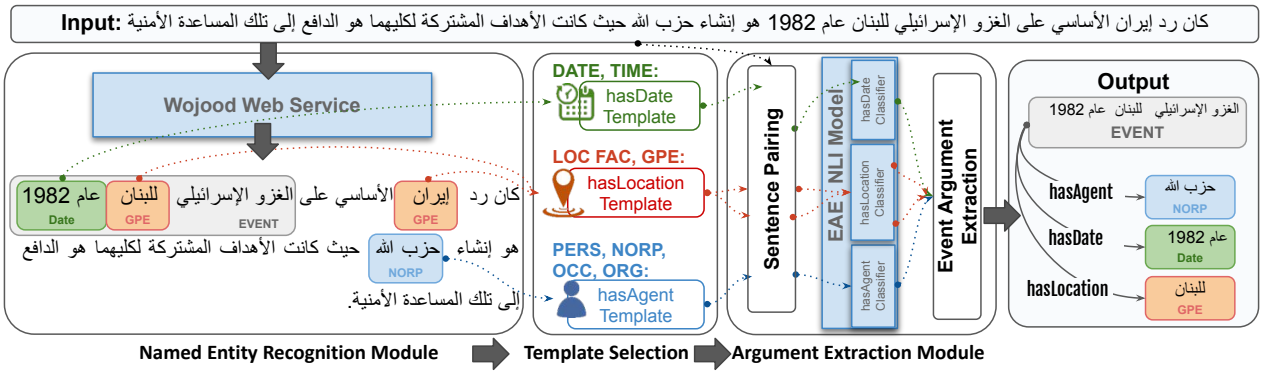


Figure 6: End-To-End Event Argument Extraction Architecture.

Class	Support	P	R	F_1
Positive	778	90.06	92.42	92.24
Negative	1435	95.99	95.68	95.78
Average				**94.01**

Table 6: Results on the ***Hadath***$^{\text{NLI}}$ Test Set.

Relation	Support	P	R	F_1
hasAgent	111	82.60	85.58	84.07
hasLocation	267	92.40	85.70	89.88
hasDate	403	95.38	97.51	96.44

Table 7: Results on the ***Hadath***$^{\text{NLI}}$ Test Set, per relation

Relation	Count
hasAgent	138
hasLocation	218
hasDate	125
Total	**481**

Table 8: Number of relations in ***Wojood***$^{\text{OutOfDomain}}$

Pairs	Positive	Negative	Total
hasAgent	108	304	**412**
hasLocation	201	207	**408**
hasDate	124	180	**304**
Total	**433**	**691**	**1124**

Table 9: NLI datasets based on ***Wojood***$^{\text{OutOfDomain}}$

Class	Support	P	R	$F1$
Positive	478	71.05	78.03	74.38
Negative	1809	94.04	91.60	92.80
Average				**83.59**

Table 10: Experimental Results on ***Wojood***$^{\text{OutOfDomain}}$.

the system evaluates the extracted entities. If an event entity is recognized in a sentence, then other entities in this sentence are considered candidate arguments for this event.

Template Selection Module: For each entity, a template is selected based on its category and used to construct the hypothesis.

Argument Extraction Module: The input sentence is paired with the template and sent to the EAE NLI model to identify the argument entities for an event and their corresponding relationships. The type of template serves as the basis for establishing the relationship between the event and the entity. Specifically, if the EAE NLI model indicates a positive connection between the input sentence and the template, then the entity and the event linked within the template are considered to have a relationship indicated by the template label.

9.2 Computing EAE baselines

The performance of our proposed EAE system is evaluated using ***Hadath***$^{\text{NLI}}$ and ***Wojood***$^{\text{OutOfDomain}}$ test sets. Results are shown in Table 11. Note that the relation classifiers share parameters, and the evaluation assumes named entities to be correctly recognized.

Dataset	P	R	F_1
Hadath$^{\text{NLI}}$	93.45	94.52	93.99
Wojood$^{\text{OutOfDomain}}$	67.79	83.68	74.90

Table 11: Baselines: evaluation of our EAE system.

10 Ablation Studeis

Best Template: to choose the best template to implement in the EAE system for each relation (Figure 5), we evaluated them using the test set. Table 12 shows that t_2 performed slightly better.

Template	$P.$	$R.$	$F1.$
t_1	92.43	92.65	92.54
t_2	92.63	92.60	**92.61**
t_3	92.01	92.83	92.40
t_4	92.35	92.40	92.37

Table 12: Ablation study to choose the best template.

Best loss function: We compared two loss functions: cross-entropy loss (L_{CE}) and Noise Contrastive Estimation Loss (L_{NCE}). Table 13 shows that while L_{NCE} slightly outperforms L_{CE} on the ***Hadath***$^{\text{NLI}}$, its performance significantly improves on ***Wojood***$^{\text{OutOfDomain}}$, showing its efficacy across diverse contexts.

Additionally, we improved the model's performance by combining weighted cross-entropy loss (L_{WCE}) with L_{NCE}, using weights $w_p = 1$ and $w_n = 0.5$. This approach, with higher weights for positive relations and slightly lower weights for negative ones, yielded the best results.

Loss Fn.	**Class**	**Support**	P	R	$F1$
Cross	Neg.	1494	94.94	87.18	90.89
Entropy	Pos.	387	62.94	82.43	71.38
(L_{CE})	Avg.				81.14
$Loss$	Neg.	1494	95.88	90.29	93.00
($w_p = 1$,	Pos.	387	69.41	85.01	76.42
$w_n = 1$)	Avg.				84.71
$Loss$	Neg.	1494	94.44	92.10	93.26
($w_p = 1$,	Pos.	387	72.17	79.07	75.46
$w_n = .2$)	Avg.				84.36
$Loss$	Neg.	1494	95.75	90.56	93.09
($w_p = 1$,	Pos.	387	69.87	84.50	76.49
$w_n = .5$)	Avg.				**84.79**

Table 13: Results (F1-score %) on ***Hadath***$^{\text{NLI}}$ test set.

11 Conclusion and Future work

Our introduction of the ***Wojood***^{Hadath} and ***Wojood***$^{\text{OutOfDomain}}$ corpora significantly advances Arabic event-argument extraction by providing a rich dataset with high inter-annotator agreement. Our novel BERT-based method for event relation extraction demonstrates exceptional performance, achieving high F_1 scores on both the ***Hadath***$^{\text{NLI}}$ dataset and on ***Wojood***$^{\text{OutOfDomain}}$ dataset. Additionally, our implementation of the EAE end-to-end system as part of the open-source SinaTools will enrich the Arabic NLP industry.

Large language models (LLMs) can further enhance our work by extracting information, including named entities and relationships, from text to populate knowledge graphs and improve other knowledge graph tasks like embedding and completion (Barbon Junior et al., 2024). In future work, we will explore the capabilities of LLMs to enhance event-argument extraction. Integrating LLMs into our framework could potentially improve the accuracy and scalability of our event extraction system.

Limitations

The constructed ***Wojood***^{Hadath} and ***Wojood***$^{\text{OutOfDomain}}$ corpora primarily focus on MSA data and do not cover dialectal variations. Furthermore, even though we included out-of-domain tests to assess performance, our results are constrained to the specific domains used in our study.

Ethics Statement

The corpora provided for this research are derived from public sources, eliminating specific privacy concerns. The results of our research will be made publicly available to enable the research community to build upon them for the public good and peaceful purposes. Our data, tools, and ideas are strictly intended for non-malicious, peaceful, and non-military purposes.

Acknowledgements

This research is partially funded by the research committee at Birzeit University. We extend our gratitude to Taymaa Hammouda for the technical support and to the students who helped and supported us during the annotation process, especially Haneen Liqreina and Sana Ghanim.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Kashif Ahmad, Konstantin Pogorelov, Michael Riegler, Nicola Conci, and Pål Halvorsen. 2019. Social media and satellites: Disaster event detection, linking

and summarization. *Multimedia Tools and Applications*, 78:2837–2875.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. Arab-GlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.

Mohammad AL-Smadi and Omar Qawasmeh. 2016. Knowledge-based approach for event extraction from arabic tweets. *International Journal of Advanced Computer Science and Applications*, 7(6).

Ebtesam Alomari, Iyad Katib, and Rashid Mehmood. 2020. Iktishaf: A big data road-traffic event detection tool using twitter and spark machine learning. *Mobile Networks and Applications*, pages 1–16.

Sylvio Barbon Junior, Paolo Ceravolo, Sven Groppe, Mustafa Jarrar, Samira Maghool, Florence Sèdes, Soror Sahri, and Maurice Van Keulen. 2024. Are Large Language Models the New Interface for Data Pipelines? In *Proceedings of the International Workshop on Big Data in Emergent Distributed Environments*, BiDEDE '24, New York, NY, USA. Association for Computing Machinery.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. Red^{FM}:a filtered and multilingual relation extraction dataset. *arXiv preprint arXiv:2306.09802*.

Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Zero-shot cross-lingual event argument extraction with language-oriented prefix-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12589–12597.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2018. A tf-idf and co-occurrence based approach for events extraction from arabic news corpus. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 272–280. Springer.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A Panoramic survey of Natural Language Processing in the Arab Worlds. *Commun. ACM*, 64(4):72–81.

Lina Duaibes, Areej Jaber, Mustafa Jarrar, Ahmad Qadi, and Mais Qandeel. 2024. Sina at FigNews 2024: Multilingual Datasets Annotated with Bias and Propaganda. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Niama El Khbir, Nadi Tomeh, and Thierry Charnois. 2022. ArabIE: Joint entity, relation and event extraction for Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 331–345, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + Baladi: Towards a Levantine Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Tymaa Hammouda, Mustafa Jarrar, and Mohammed Khalilia. 2024. SinaTools: Open Source Toolkit for Arabic Natural Language Understanding. In *Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024)*, Procedia Computer Science, Dubai. ELSEVIER.

Aldo Hernandez-Suarez, Gabriel Sanchez-Perez, Karina Toscano-Medina, Hector Perez-Meana, Jose Portillo-Portillo, Victor Sanchez, and Luis Javier García Villalba. 2019. Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. *Sensors*, 19(7):1746.

Emna Hkiri, Souheyl Mallat, and Mounir Zrigui. 2016. Events automatic extraction from arabic texts. *International Journal of Information Retrieval Research (IJIRR)*, 6(1):36–51.

Mustafa Jarrar. 2021. The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. *Applied Ontology Journal*, 16(1):1–26.

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa' Omar. 2023a. WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 748–758. ACL.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic-multilingual database with a lexicographic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer.

Mustafa Jarrar and Werner Ceusters. 2017. Classifying processes and basic formal ontology. In *Proceedings of the 8th International Conference on Biomedical Ontology (ICBO 2017)*, volume 2137. CEUR Workshop Proceedings.

Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022a. Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022b. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3626–3636.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023b. SALMA: Arabic Sense-annotated Corpus and WSD Benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 359–369. ACL.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Nouran Khallaf, Elin Arfon, Mo El-Haj, Jon Morris, Dawn Knight, Paul Rayson, and Tymaa Hammoudaand Mustafa Jarrar. 2023. Open-source thesaurus development for under-resourced languages: a welsh case study. In *The 4th Conference on Language, Data and Knowledge (LDK2023)*.

Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, and Muhammad Abdul-Mageed. 2023. Arabic Fine-Grained Entity Recognition. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 310–323. ACL.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *ArXiv*, abs/2010.04592.

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955.

Tomer Simon, Avishay Goldberg, and Bruria Adini. 2015. Socializing in emergencies—a review of the use of social media in emergency situations. *International journal of information management*, 35(5):609–619.

Sonit Singh. 2018. Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*.

Jiashu Xu, Mingyu Derek Ma, and Muhao Chen. 2022. Can nli provide proper indirect supervision for low-resource biomedical relation extraction? *arXiv preprint arXiv:2212.10784*.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. *CoRR*, abs/2210.12714.