IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 April 2023, accepted 22 May 2023, date of publication 31 May 2023, date of current version 7 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3281680

RESEARCH ARTICLE

A Parallel Corpus-Based Approach to the Crime Event Extraction for Low-Resource Languages

NINA KHAIROVA[1,2], ORKEN MAMYRBAYEV[3], NINA RIZUN[4], MARIIA RAZNO[5], AND YBYTAYEVA GALIYA[6]

[1]Department of Intelligent Computer Systems, National Technical University "Kharkiv Polytechnic Institute," 61002 Kharkiv, Ukraine
[2]Department of Computer Science, Umeå University, 901 87 Umeå, Sweden
[3]Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan
[4]Department of Informatics in Management, Gdańsk University of Technology, 80-233 Gdańsk, Poland
[5]Institut für Slawistik und Kaukasusstudien, Friedrich Schiller University Jena, 07743 Jena, Germany
[6]Department of Cybersecurity, Information Processing and Storage, Satbayev University, Almaty 050013, Kazakhstan

Corresponding author: Nina Khairova (khairova.nina@gmail.com)

This work was supported by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan under Grant AP09259309. The information model and software of automatic search and analysis system of multilingual illegal web content based on an ontological approach.

ABSTRACT These days, a lot of crime-related events take place all over the world. Most of them are reported in news portals and social media. Crime-related event extraction from the published texts can allow monitoring, analysis, and comparison of police or criminal activities in different countries or regions. Existing approaches to event extraction mainly suggest processing texts in English, French, Chinese, and some other resource-rich and well-annotated languages. This paper presents a parallel corpus-based approach that follows a closed-domain event extraction methodology to event extraction from web news articles in low-resource languages. To identify the event, its arguments, and the arguments' roles in the source-language part of the corpus we utilize an enhanced pattern-based method that involves the multilingual synonyms dictionary with knowledge about crime-related concepts and logic-linguistic equations. The event extraction from the target-language part of the corpus uses a cross-lingual crime-related event extraction transfer technique that is based on supplementary knowledge about the semantic similarity patterns of the considered pair of languages. The presented approach does not require a preliminarily annotated corpus for training making it more attractive to low-resource languages and allows extracting TRANSFER, CRIME, and POLICE types of events and their seven subtypes from various topics of news articles simultaneously. Implementation of our approach for the Russian-Kazakh parallel corpus of news portals articles allowed obtaining the F1-measure of crime-related event extraction of over 82% for the source language and 63% for the target language.

INDEX TERMS Cross-lingual transfer, crime analysis, event extraction, low-resource language, natural language processing, parallel corpus, semantic annotation.

I. INTRODUCTION

Event extraction (EE) is an important and contemporary task in NLP, which is a part of information extraction (IE). It generally focuses on the mining of particular events in unstructured text and the determination of the types and attributes of these events.

Today there exist a lot of different techniques for event extraction. Most of them are domain-specific and focus on a small number of relations in specific preselected domains [1], [2], [3]. These approaches are usually named closed-domain event extraction (CdEE). The more recent IE systems use a domain-independent architecture and a sentence analyzer, which are the so-called open-domain event extraction (OdEE) systems. Such applications operate without predefined event schemas, and the extraction aims at detecting events from a sentence or phrase and clustering

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

similar events via extracted event keywords [5]. However, despite the different approaches, EE remains a very challenging task for text processing, especially for low-resource languages.

Furthermore, today the problem of developing software tools and mechanisms to support the investigation and prevention of criminal acts based on textual information distributed on the Internet remains unresolved. There are a lot of applications aimed at addressing the issue, including hate speech detection, crime pattern modeling, crime prediction, crime-related topic identification, etc. [5], [6], [7]. Nevertheless, in order to prevent and investigate crimes, it is still necessary to use more efficient methods to analyze the criminal content in computer-mediated communication (CMC). One of the most challenging and efficient solutions for the identification and prediction of crime actions is crime-related event extraction (CREE) [8], [9]. Opportunities to obtain structured information about an occurred or prospective crime, accumulate this information, and then model a crime pattern, are the cause of the CREE task effectiveness. Meanwhile, despite the existence of the studies related to the CREE task, the extraction of various types of crime-related events (CRE) from texts, which are not police or witness narrative reports, remains a big challenge. Additionally, the Event Extraction task becomes more difficult for texts written in low-resourced and under-annotated languages.

This study addresses the crime-related event extraction from web news articles. The proposed parallel corpus-based (PaCo-based) approach (i) follows the CdEE methodology; and (ii) utilizes such linguistic resources as an aligned corpus [10] and a multilingual synonyms dictionary with additional knowledge about criminal-related concepts [11]. Proposed approach allows us to solve the following problems: (1) search and extract a wide range of crime-related and police-related events such as Traffic Accident, Injure, Trials, Police Activities, and others, simultaneously; (2) extract these events directly from web news articles instead of having to search for access to specific police crime reports or witness narrative reports; and (3) extract these events from articles in low-resource and under-annotated languages.

Method-wise, a PaCo-based approach is *two-fold* and uses:

(1) *Enhanced pattern based* (EPB) method for CREE from the first part of the corpus (source language); and

(2) *Cross-lingual CRE transfer* technique for processing the second part of the corpus (target language), based on supplementary knowledge about the semantic similarity patterns of the considered pair of languages.

The method of EPB event extraction is based on the definition of an event provided by the Guide to annotating events for automatic content extraction (ACE) 2005 [12]. According to it, an event is a specific occurrence of something that happens at a certain time and a certain place involving one or more participants. In our study, we consider only CRE in the close-domain area; that is, we extract events concerning a crime or unlawful actions. The proposed EPB event extraction method is close to the pattern matching technique [13], which first constructs some specific event templates, and then the templates match with raw text or annotated text. However, generally, EE applications that are based on this approach allow exploiting relatively few patterns for EE. To create more CRE patterns, we develop and enhance the pattern matching technique using the logic-linguistic model [14] that allows representing participants and attributes' roles of the event via the relations between grammatical and semantic features of the words in a sentence or phrase. This approach provides an opportunity to describe every possible attribute role that exists in a particular language and a particular domain.

The introduced cross-lingual CRE transfer technique that employs for conveying events and their arguments labels from the source language into the target language of the corpus, is based on a hypothesis that the same meaning of aligned sentences exists in both languages' parts of the parallel corpus. To extract CRE from the target language part of the corpus, we simultaneously use the (1) preliminary POS-tag labeling of target language texts; and (2) the patterns of the correspondence between POS-tags of target language sentences and possible roles of the event participants/attributes that are transferred from an aligned source language sentence.

Thus, the main *methodological*contribution of this paper is the introduction of the two-stage approach to extract criminal and police-related events from a bilingual parallel corpus covering two low-resource and under-annotated languages. The *methodological* contribution of *EPB method* part of our approach is: (1) identifying patterns of three main types and seven subtypes of events related to crime and police activities; (2) expanding the list of crime and police-related types of events that can be extracted from news articles; as well as the (3) proposal of logical-linguistic equations that allow determining the roles of participants and the attributes of events through the relationship between the semantics and characteristics of words in sentences in the language under consideration. The *methodological* contribution of the *cross-lingual CRE transfer technique* part of our approach lies in the experimental proof of the possibility of obtaining significant results from the application of the method of using preliminary labeled events in one language texts to label events in another language texts of a parallel aligned corpus. Thus, our approach does not require a pre-annotated learning corpus, which is especially important for low-resource languages.

Furthermore, the proposed approach can serve as a supplementary tool for *automatically designing event-labeled corpora*. In practice, creating a training corpus with event annotations typically involves identifying a set of relevant events and manually annotating a large corpus of text, or manually labeling a smaller subset of the corpus to train a model, and then applying it to the larger unlabeled corpus to identify additional instances of the target events [23], [37]. However, both of these approaches can be time-consuming and expensive, particularly when dealing with large amounts

of text that require annotation [22]. Our approach enables automatic identification of event types and attributes in the initial labeling stage, with subsequent manual adjustments, which significantly reduces the cost of implementation and reproduction.

The main *practical* contribution of our research is to obtain (i) the distribution of events by subtypes related to crime and police activities, extracted from the corpus of Russian and Kazakh news articles; and (ii) sample for POS tags of Kazakh text fragments, which may represent various types of CRE structure.

The rest of the paper is structured as follows. Section II gives an overview of the articles corresponding to the existing approaches to automatic Event Extraction, examines text mining applications in crime, and views corpora that are utilized to evaluate and refinement of EE methods. Section III introduces the types and subtypes of crime-related events identified and extracted in the study. Section IV presents the methodology for crime-related events extraction from two parts of the corpus (source language part and target language part) separately. Section V describes a qualitative evaluation of the proposed approach, whereas Section VI discusses the scientific and practical contributions of the research, its limitations, and future work. A demonstration of crime-related events extracted by our approach are available at the GitLab repository.[1]

II. RELATED WORK

The main purpose of this section is (1) to gain an understanding of the current knowledge base and underlying trends with respect to three main aspects: (i) automatic event extraction as a method of study; (ii) crime-related text analysis as a chosen problem area; and (iii) Linguistic resources-based solutions as an approach to increase the power of the event extraction method; and (2) to identify major gaps in the state of the art in crime-related Events extraction in the context of low-resourced and under-annotated languages.

A. METHODS OF EVENT EXTRACTION FROM TEXTS

The existing methods related to EE tasks could be categorized into four groups, namely, (1) pattern matching algorithms, (2) machine learning methods, (3) deep learning models, and (4) unsupervised machine learning methods.

The first group exploits *pattern-based* event extraction approaches. Such an approach was first proposed in 1993 by Riloff [14] to extract terrorist events from domain-specific texts. Now there are quite a lot of pattern-based EE systems which are domain-specific for extracting various types of events [3], [15]. The recent trends of studies of EE in the terrorism and criminal domain are of considerable interest to our research. José A. Reyes-Ortiz [3] presented an approach based on pattern matching techniques to extract criminal events from Spanish texts. To evaluate the process results, the author used a set of manually tagged newspapers with categories of specific events. Thus, Li et al. [16] applied EE technology to the case description part in the Chinese legal text. The authors defined the event type, event arguments, and event arguments roles of the larceny case. Abdelkoui and Kholladi [17] described the EE of criminal incidents from Arabic tweets. The author's approach is based on combining various indicators, including the names of places and temporal expressions that appear in the tweet messages.

However, the most recent papers devoted to EE tasks belong to the second group of approaches based on *machine-learning* (ML). These approaches apply traditional ML classification algorithms, like support vector machine (SVM), maximum entropy (ME), the nearest neighbor, and others. Most commonly, these algorithms utilize POS tags, lemmatized words, the type of syntactic dependency between a trigger word and entity, and the dependent words and entity types as features of event classifications. More often, EE approaches based on ML algorithms were utilized in domain-specific areas, for instance, in the biomedical domain or finance and economic-connected domains. Some authors suppose simultaneously using pattern-based approaches and models of ML or deep ML for EE. The paper [18] proposed a regularization-based pattern balancing method (RBPB) that includes using both event patterns in a sentence to identify an event and the SVM classifier to define the trigger type. At the same time, these state-of-the-art systems, based on traditional ML methods, require many complex and hand-designed features. To generate these features, it is necessary to have professionals with linguistic knowledge and experts with domain knowledge. Additionally, these features often are represented by one-hot vectors, which cause data sparsity and feature selection problems [19].

On the other hand, according to Christopher Manning [20], *deep learning* techniques can be successfully applied in various NLP tasks related to classification, particularly with the classification of sentences, words, or full texts. Consequently, since the task of EE is related to the sentences and words classification issue, we can expect progress in applying deep learning techniques for extracting events from texts in the nearest future. Using convolutional neural networks (CNN) and recurrent neural network models in EE in the last few years is illustrative in this regard. For example, Yagcioglu et al. [8] employed CNN and a long short-term memory (LSTM) recurrent neural network to detect cyber security events from a noisy short text. The graph neural networks (GNN) use multiple neurons operating on a graph structure to enable deep learning in non-Euclidean spaces. Thus, in [21], the authors proposed to jointly extract multiple event triggers and arguments by attention-based graph convolutional networks. Liu et al. [22] have applied attention mechanisms in neural models in order to guide a neural model to unequally treat each component of the input according to its importance to the EE task. However, today usage of machine learning and deep learning is still great challenges in practice. The main reason is the need to handle a large, annotated corpus for model training. Usually, obtaining such

[1] Parallel corpus.

a corpus is a time-consuming and labor-intensive task, which involves a lot of domain and professional experts.

To avoid the necessity of the labeled corpus, some scholars leveraged *unsupervised learning* approaches. In these cases, they focused on open-domain EE tasks [4]. Open-domain EE approaches operate without predefined event schemas, and usually, this extraction aims at detecting events from a sentence or phrase and clustering similar events via extracted event keywords. However, in the case of the open-domain EE approaches, the accuracy of EE turns out to be rather low, and the events themselves are mostly vague and blurred.

While the EE has become a mature academic field, this task becomes challenging for the texts written in low-resourced and under-annotated languages. For that reason, in the last few years, studies addressing cross-lingual learning (CLL) for EE appeared [23]. Over the past decade, we have also seen exploitation of the multilingual BERT model and CNN [24] for cross-lingual relation and event extraction. Meanwhile, in most cases, cross-lingual event extraction approaches were based on multilingual versions of the ML models pre-trained on large multilingual corpora [25], and resource-rich and well-annotated language were exploited as the source language of the corpus. Typically, that was English [23], [24]. Moreover, EE requires a rich label space. That is an additional reason why gold-standard annotations for event extraction are publicly available only for a few languages [25]. To fill this gap, when there are no well-annotated corpora for specific languages, we suppose that it may be possible to employ supplementary knowledge about the similarity of the syntactic and semantic patterns of the considered pair of languages to Cross-lingual EE transfer [24].

B. TEXT MINING APPLICATIONS IN CRIME

Over the past several years, the number of research related to crime has grown significantly. To comprehensively include various existing tasks related to crime text information, we propose the following classification of Text Mining applications in crime: (1) crime texts identification (or crime detection); (2) crime event types classification; (3) crime pattern modeling and crime prediction; (4) hate speech detection; (5) crime information extraction (CIE), including crime entities (CE) identification; (6) crime-related event extraction. Even though sometimes these directions can overlap, Appendix D shows the generalized information about the existing approaches.

Perhaps most of the current studies relevant to the problems of crime-related texts analysis, address the selection of such kinds of texts. The articles on crime text *identification* commonly described traditional clustering and topic identification approaches [25].

Nevertheless, much research not only distinguished between crime-related news/information and not-crime-related news/information but focused on the *classification* of the crime event types. Salas [7] selected two algorithms (support vector machines and neural networks) to multi-classify what type of crime news is reported. They processed crime articles from the Spanish corpus of Peruvian news. Meanwhile, in the recent overview, Hassant et al. [5] deduced that the most popular techniques typically chosen in the different applications for the crime event types classification are SVM and Neural Networks algorithms. Besides the classification of the crime-related texts, we can also distinguish studies related to crime *pattern modeling* and crime prediction, which are often based on additional police-recorded crime data attributes [26].

Such a direction of research related to crime text information as *hate speech detection* in social media texts [27], [28], [29] should be highlighted separately. Generally, the authors considered this term for numerous kinds of insulting user-created content on Twitter, blogs, and other social networks. In the broader sense, the term Hate Speech refers to any communication that disparages a person or a group based on some characteristic, such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or others [6]. In general, the problem of hate speech detection is solved by supervised machine learning classifiers [29] or, more often, by using recurrent or convolutional neural networks [28].

Over the past few years, many papers dedicated to the issue of *crime information extraction* appeared, which considered the information extraction task about occurred or prospective crimes. Usually, either crime police reports or open sources of textual information were utilized as a dataset for the CIE task. However, obviously the best results were achieved when information was extracted from crime reports. Das and Das [30] dealt with crime reports for the USA, UAE, and India. The authors demonstrated a graph-based clustering to extract paraphrases from the crime dataset for subsequent labeling of crime reports. Other studies considered particular types of criminal offenses based on textual information from open sources. Dasgupta et al. [31] leveraged computational linguistics-based methods to extract different crime-related entities and events from crime-related news documents. They extracted the name of the criminal, the name of the victim, the nature of the crime, the geographic location, date and time, and the action taken against the criminal by using probabilistic classifiers and domain ontology to augment the accuracies of the extraction process. Rare research combined the use of police reports, newspaper articles, and victims' and witnesses' crime narratives.

The task of crime information extraction also includes *crime entities extraction* [9], [32]. Joseph et al. [9] gained information about places, most used drugs, amount of each drug from reported news. They processed using NLP techniques like NER for extracting structured information. At the same time, Das & Das [32] extracted classes named entities such as states, streets, towns/cities, villages, and male forenames from online newspapers and websites that provide reports about crimes against women sorted according to the states.

Despite the large number of studies concerning crime-related texts generally, we can't say that there are many papers regarding the *crime-related event extraction* task. Although, the task of CREE from natural language texts had first arisen at the early DARPA Message Understanding Conferences (MUCs). The domain of MUC-3 and MUC-4 was Latin-American Terrorism, and the events extracted were associated with particular terrorist actions. In contemporary studies, crime-related events are often defined as various types of events that refer to criminal activities. Typically, research studies consider the problem of CREE separately for various types of events (related to terrorism, cybercrime, crimes against the person, crimes related to transport, etc.) From this perspective, in particular, the works [9], [33] aimed to extract available drug crime and substance abuse information from online newspaper articles. Rahem and Omar [33] obtained information about the nationalities of drug dealers, names of drugs, and the quantity and prices of drugs in the local market. Their extraction system was based on grammatical and heuristic rules and data from Malaysian National News Agency (BERNAMA). In the paper by Yagcioglu et al. [8], cybersecurity events detection was considered. Authors focused on such cybersecurity events as zero-day exploits, ransomware, data leaks, security breaches, vulnerabilities, etc. For their approach, they utilized a manually labeled dataset that included 2K tweets about crimes against women in India. Meanwhile, the study of Hossain et al. [25] aimed to predict violent events, such as military action by state actors or terrorist attacks by non-state actors (MANSA events). For evaluation of the approach, they used manually extracted, structured reports on events at the actor, city, and country levels.

Several studies performed the extraction of events connected exactly with a hate crime. According to the FBI's UCR Program1, a hate crime can be a criminal offense against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity. For instance, Davani et al. [34] provided event detection and event extraction from news articles based on a crime acts taxonomy. Authors considered homicide and kidnapping events and such event attributes as the target of a crime event and the type of crime. For experiments, they manually annotated subsets of the main unlabeled local news articles corpus.

Despite the widespread development of approaches used for text mining of crime-related texts, the solutions presented in the extant literature are mainly based on the ML and Deep ML models [28], [29], [35]. Such models, as we mentioned in the section II-B, require presence of large corpora, which should be previously balanced and manually tagged by experts under clear rules, and provided with language subtleties [29].

C. LINGUISTIC RESOURCES-BASED SOLUTIONS

Following the conclusions made in the sections II-A and II-B, it can be argued that for the evaluation and refinement of EE methods *corpora* are frequently utilized. These corpora should be specially annotated by semantic labels, which may describe event types, for instance, Socio-political events (SPE) and event arguments such as a person, organization, location, time, geopolitical entity, facility, vehicle, weapon, and others. Thus, the DEFT richer event description annotation corpus, developed by the Linguistic Data Consortium, includes 158 documents as a prior training set and 202 additional documents as a test set [36]. Now the corpus annotation scheme comprises 8731 events and 10319 entities and can be utilized to formally evaluate approaches to EE tasks from English, Chinese and Spanish news articles and discussion forums.

Event-annotated corpora are most often focused on specific problem domains. So, one of the most developed are corpora of biomedical information. Ramponi et al. [37] analyzed some public resources that provide manually annotated events in the biomedical field, including the GENIA event corpus, the BioInfer (biomedical information extraction resource) corpus, the gene regulation event corpus (GREC), the GeneReg corpus, and some others. Over the past few years, the use of linguistic resources for the study of crime-related topics has also intensified [38]. The use of ontologies, corpora, thesauri, and structured lexical bases is still the most relevant for the hate speech detection task. Paper by Çöltekin [39] is centered on the notion of hate speech to date and introduced the first corpus of Turkish offensive language that consists of randomly sampled micro-blog posts from Twitter. A paper by Kumar et al. [40] concerned the problem of the annotated corpus creation based on Hindi-English code-mixed data of Twitter and Facebook. The corpus is annotated using an aggression tag set. In the latest study by Battistelli et al. [41], the methodology to build an ontology of the online hate speech domain in French was presented, but at the same time, unfortunately, the paper focused on modeling development aspects, while practical using of the ontology to annotate texts was not addressed. Furthermore, the popularity of that research field can be confirmed by it provided in SemEval-2020 [42] tasks as Task 6: Identifying and categorizing offensive language in social media (OffensEval) and Task 12: Multilingual offensive language identification in social media (OffensEval 2020) accordingly. To estimate different approaches to offensive language identification, automatic categorization of offense types, and offense target identification, the tweets collections in English, Arabic, Danish, English, Greek, and Turkish [42] were annotated according to the hierarchical taxonomy of the OLID schema were utilized. Apparently, there is a well-structured hierarchical system for detecting hate speech. However, there is no such general scheme for CREE yet.

At the same time, there are quite a lot of corpora focused on a subgenre of legal and judicial texts (the Cambridge Corpus of Legal English, The House of Lords Judgments Corpus, The Proceedings of the Old Bailey, JUD-GENTT, A Corpus of Malawi Criminal Cases) [43]. In many cases, text mining tasks related to crime were based on the corpora

of newspaper articles. Mukherjee and Sarkar [44] proposed to exploit the corpus of newspapers written in the Bengali language to automatically get a picture of high crime-prone locations. Adily et al. [45] utilized the corpus of 492,393 domestic violence events provided by the New South Wales Police Force [46].

In addition, special mention should be made of the most current CREE approaches, which are based not only on the annotated corpora, but also *ontologies* (or *specialized lexicons*). Thus, de Mendonça et al. [47] proposed the ontology-based framework for criminal intention classification (OFCIC). They employed the ontology of criminal expressions (OntoCexp) to select potentially crime-related posts on Twitter.

Thus, we can *summarize* the results of our related work analysis as follows:

(1) although the field of Event Extraction (EE) has shown significant progress as a scientific and practical direction, it remains challenging to tackle when dealing with texts written in low-resourced and under-annotated languages. Many existing EE approaches, including those focused on crime-related domains, rely on multilingual ML or Deep Machine Learning models that are pre-trained on large corpora in well-annotated languages. To address this limitation, our study proposes enhancing the Cross-lingual EE approach by incorporating additional knowledge about the similarity of syntactic and semantic patterns between the languages being considered for transfer. This enriched approach aims to bridge the gap in tackling EE tasks in low-resourced and under-annotated languages;

(2) the development of linguistic resources, such as corpora, is often employed for evaluating and refining Event Extraction (EE) methods. These resources are typically created for highly specialized problem domains, and there is a growing focus on utilizing them for low-resource and under-annotated languages to enhance the effectiveness of CRE approaches in a multilingual context.

III. TYPES AND SUBTYPES OF CRIME-RELATED EVENTS

Following the studies [9], [33], [34], [35], we determine and extract CRE from a corpus of news articles related to police and criminal activities. However, unlike previous research, despite the limited count of the event types, we consider not specific types of crimes (only drug crime or only traffic incidents, etc.), but the big group of events that relates to unlawful action (Traffic Accident, Hate crime, Police Activities, and others).

Specifically, we are interested in TRANSFER, CRIME, and POLICE types of events and their seven subtypes. Table 1 shows event types and subtypes considered.

Generally, in all these kinds of CRE, we can say about two participants and several attributes of the action or event. The Agent is a participant that is an initiator of an event. The second participant of these event types is an Object which, in a general way, is represented by a person, an organization, or a vehicle, to which the event action is directed.

TABLE 1. Event types and subtypes that we consider.

#	Event type	Event subtype
1	TRANSFER	Movement, Traffic Accident
2	CRIME	Injure, Offense
3	POLICE	Arrest, Trial, PD

Based on the Coplink project [48], to determine participants of CRE, we distinguish three different types of entities that can be involved in a criminal action. We employ semantic classes of people names, organizations names, and vehicles. However, various types and subtypes of CRE can involve various entity types in their capacity as Agent and Object. Additionally, all the types and subtypes of events we are considering, have traditional TIME-ARG and PLACE-ARG attributes. Sometimes we look for the Instrument or device to determinate modus operandi, for example, a weapon applied to inflict harm. Extra, on rare occasions, we can use an optional slot WHY-ARG to describe the reason for the event.

A CRIME CRE occurs whenever a person or an organization does something criminalized or unlawful. There are two subtypes of a CRIME event: INJURE and OFFENSE. An INJURE subtype of a CRIME CRE occurs whenever an action covers a person entity, so-called crimes against persons. This person can experience physical harm (be killed, be injured) or be affected by other criminal actions (be robbed, be tricked). Consequently, an Object can be only the harmed person(s), whereas an Agent of the subtype is the initiator of the attacking action, a person or an organization damaging to the physical harm.

An OFFENSE subtype occurs whenever an object of the criminal action isn't a person directly. In this case, an OFFENSE event can have two or one participant and some attributes of the event. The agent is the initiator of the offense action, a person or an organization damaging to some harm or doing an illegal activity. It is a necessary participant in the event. Nevertheless, an inanimate OBJECT, which is the second participant of this subtype, can either be or not in a certain phrase or sentence.

A TRANSFER CRE includes two subtypes, namely, MOVEMENT and TRAFFIC ACCIDENT. A MOVEMENT subtype of a TRANSFER Event occurs whenever an inanimate object or a PERSON is moved from one LOCATION to another. At the same time, we have suggested that moving something to steal or thieve is not a MOVEMENT CRE, it is exactly a CRIME CRE. Another subtype of a TRANSFER CRE is a TRAFFIC ACCIDENT, which occurs whenever a vehicle suffers an accident. In this case, an Agent should be a person or a vehicle that triggered the accident.

The last type of event that we have considered as CRE is a POLICE Event that occurs whenever the action is going to be done by police or officials. A POLICE CRE includes three subtypes, namely, ARREST, TRIAL, and PD. An ARREST is a subtype of a POLICE CRE, which occurs whenever the movement of a person is going to be constrained by a state

actor (for instance, policemen or justice). In the case of an ARREST subtype, Agent can be well-defined as a person or an organization that was an initiator of the detention of another person, whereas an Object is only a detained person.

A TRIAL is a subtype of a POLICE CRE, which occurs whenever a court or some government organization accuses a person or an organization of committing a crime. A PD (Police Department) is a subtype of a POLICE CRE, which occurs whenever a police officer implements official duties. The Agent of a PD subtype should be exactly a policeman as a person or a police department as an organization.

IV. METHODOLOGY

Following the previous studies [8], [9], [33], [34], we determine and extract three types and seven subtypes of CRE from the corpus of news articles relevant to police and criminal activities. An additional restriction is the use of a bilingual parallel corpus that includes aligned sentences in two low-resourced and under-annotated languages.

Our two-fold approach includes (1) the EPB method for CREE from the first part of the corpus (source language); and (2) cross-lingual CRE transfer technique for the second part of the corpus (target language). To implement this approach, in the *first* step, we use the EPB method for CREE to process texts in source language. Following the approaches of CdEE [9], [20], we sequentially determine the event trigger in the phrase describing the event, the event/trigger type, and identify the event arguments and their roles. This step involves the implementation of the following three stages:

(1.1) Application of the method of simultaneous CRE trigger detection and event/trigger type identification which is based on a multilingual synonyms dictionary with crime-related lexis [11] (for a detailed description of the method, see subsection IV-A);

(1.2) Defining a schema for each CRE subtype that is based on the CRE types and subtypes discussed in section III. The schema describes particular classes of participants involved in events of this type, such as Agents or Objects. Additionally, since we consider police or criminal activity in the website news, we are always interested in the place and time of the event. Therefore, the PLACE-ARG and TIME-ARG attributes may also be relevant to the event we are parsing;

(1.3) Developing and usage the logical-linguistic equations (LLEs), and the predefined scheme of the event subtype to extract event arguments and identify their roles. The use of LLEs provides an opportunity to describe the roles of attribute participants that exist in a particular area via relations of grammatical and semantic characteristics of the words in the sentence (see subsection IV-B for a detailed description).

In the *second* step, we apply the Cross-lingual CRE transfer technique to extract events from sentences in the target language (second part) of the corpus. The use of the technique is based on the hypothesis that the same event can be expressed by both a labeled sentence of the source language and an aligned sentence of the target language in the parallel corpus

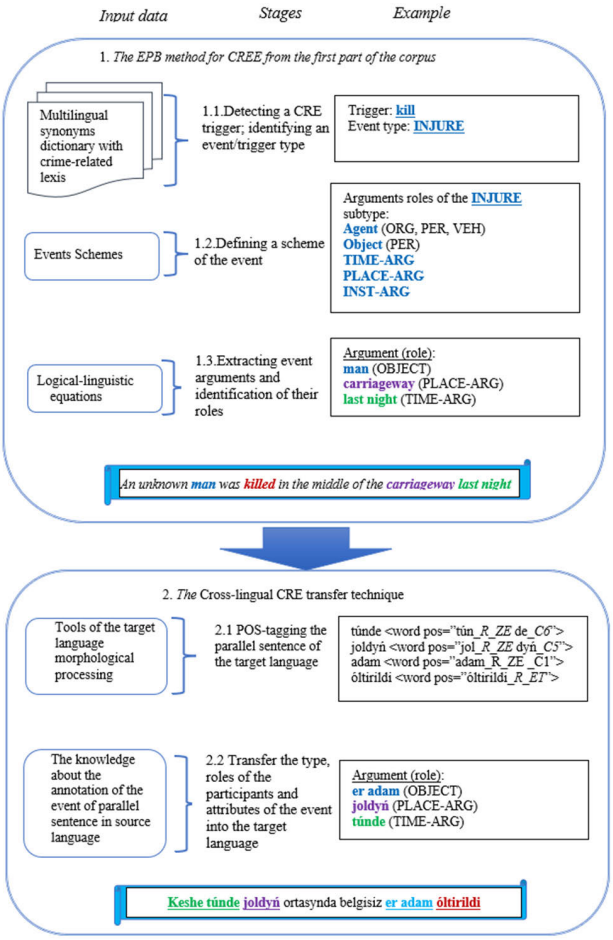


FIGURE 1. The scheme of the two-fold approach for crime-related Event Extracting from texts of a parallel corpus.

(see subsection IV-C for a detailed description). This step involves the implementation of the following two stages:

(2.1) Implementation of the POS-tag labeling of target language texts using morphological processing tools for a specific particular language;

(2.2) Using the shared semantic space of aligned sentences of the two languages, to apply knowledge about the annotation of the event of the parallel sentence in the source language to transfer the type, roles of the participants, and attributes of the event into the sentences of the target language. For this purpose, the patterns of the correspondence between POS tags of a target language sentence and the possible roles of the event participants/attributes from an aligned source language sentence are utilized. An example of applying such patterns to the Kazakh language is shown in the Subsection V.C.

Figure 1 shows the general scheme of the two-fold approach for crime-related event extracting from texts of a parallel corpus. We use the sentence "*An unknown man was killed in the middle of the carriageway last night*" in English as the source language only for making the example much clearer.

```
<vocabulary>
<nouns>
<term id="1">
<lemma lang="ru">стрельба</lemma>
<domain>OFFENSE</domain>
<synset lang="ru">обстрел, выстрел</synset>
<definition lang="ru">учебные занятия по ведению
<example lang="ru">Два человека получили ранения
<hypernims lang="ru">['приведение в действие', '
<hyponims lang="ru">['контрвыстрел', 'разряд', '
<lemma lang="en">shooting</lemma>
<synset lang="en">firing, fire, gunfire</synset>
<definition lang="en">the act of firing a projec
<example lang="en">his shooting was slow but acc
<hypernims lang="en">['actuation', 'propulsion']
<hyponims lang="en">['countershot', 'discharge',
<lemma lang="ka">атыс</lemma>
<synset lang="ka">ату, оқ жаудыру, атылыс</synse
<definition lang="ka">оқ атылғанда шығатын дыбыс
<example lang="ka">Алматының Ақбұлақ мөлтекауда
<hypernims lang="ka">['іске қосу','қозғаушы күш'
<hyponims lang="ka">['қарсы атыс', 'ату', 'басын
<lemma lang="ua">стрілянина</lemma>
<synset lang="ua">стрільба, пальба</synset>
</term>
<term id="2">
```

FIGURE 2. The fragment of the multilingual synonyms dictionary with criminal-related lexis.

A. IDENTIFICATION OF AN EVENT/TRIGGER TYPE

Our determination of a CRE trigger and identification of a trigger/event type is based on a multilingual synonyms dictionary [11]. The lexis of the dictionary is obtained manually from texts on crime-related topics. Six master's and PhD students from Kazakhstan and Ukraine were involved in the process of dictionary design for about two years. The lexis of the dictionary is based on Crime sections of Kazakhstan, Ukraine news websites, and the Texas newspaper Caller-Times corpus.

Seven main thematic categories are determined for the terms, namely Movement, Traffic Accident, Injure, Offense, Arrest, Trial, and Police Department. This choice of categories comes from the fact that the information resources, from which the texts are taken, contained most data on three criminal areas: Police, Transfer, Crime, and their subtypes mentioned above, thereby making our dictionary narrowly focused on crime-related topics.

All terms in the dictionary are separated into parts of speech, namely nouns, verbs, and adjectives. Figure 2 shows a fragment of the dictionary, which now comprises about 600 main words (325 nouns, 120 adjectives, and 170 verbs) and more than 2500 synonyms in four languages: English, Kazakh, Ukraine, and Russian. Each element <term> of the dictionary presents a word in a given part of speech with its synonyms, definitions, hyponyms, and hypernyms in four languages via child elements. A value of the elements <domain> of the dictionary indicates one of the seven aforementioned thematic categories.

Based on the statement that the main word, which most clearly expresses the occurrence of the event, is a verb [12], and consequently, a verb is the trigger of the event in a phrase or sentence, we find all verbs which occur both at the dictionary and in the texts of the first part of the corpus. The event/trigger type or the class of the event type is defined according to the value of the <DOMAIN> tag of the verb in our dictionary. For instance, the trigger verb "*kill*" in the sentence "*An unknown man was killed in the middle of the carriageway last night*" was classified as the event type "INJURE", which matches the value <DOMAIN> tag of the lemma "kill" in the dictionary.

However, because our corpus contains a kind of crime news, in some instances, some phrases or sentences describe a CRE, but they are founded on semantic light verbs, like *"mandate", "report", "assume", "give",* and some others. To take into account that kind of sentence, we consider a set of special nouns that also can be triggers of the events. We exploited the list of about 1000 nouns from our multilingual synonyms dictionary with criminal-related lexis. This list comprises, for example, such nouns as *"killer", "molestation", "gunfire", "assassination", "detonation"* and others.

B. THE USE LLEs TO IDENTIFY THE EVENT ARGUMENTS ROLES

To extract participants and attributes of the Event, we use logical-linguistic equations that identify the respective roles of the event participants according to the predefined structure of the event subtype. The main mathematical means of the LLEs is the Algebra of Finite Predicates (AFP), which allows the modeling of various finite, deterministic and discrete elements of the language system: sentences, phrases, collocations, words, grammatical and semantic characteristics, morphemes, etc.

To describe a characteristic of the language element, the AFP applies a predicate variable x_i^a, where a is a value of the characteristic of i-*th*element x [4]:

$$x_i^a = \begin{cases} 1, & \text{if } x_i = a \\ 0, & \text{if } x_i \neq a \end{cases}, \quad (1 \leq i \leq n), \tag{1}$$

where n is the amount of the elements. For example, for the Russian source language of our parallel corpus, a predicate variable x can characterize a grammatical case. In this way, x_i^{gen} will be equal to one if an i-th word of the sentence has a genitive case, while the disjunction $x_i^{\text{gen}} \vee x_i^{\text{nom}} = 1$ means that the word i can have a genitive or nominative case in the Russian sentence.

Since many grammatical and semantic characteristics of various languages are different, particular LLEs should be established for each natural language. In the pilot implementation of our approach to CREE, we consider the source language of bilingual parallel corpus capacity as the Russian language.

As possible grammatical and semantic characteristics of words in Russian sentences, representing roles of Event Arguments, we identify a grammatical case of a noun, its animate or inanimate, a semantic class of the entities, and several features formalizing the passive voice in Russian.

Thus, we introduce a finite set of six predicate variables $M = \{x, y, z, m, l, f\}$, which can characterize the words in Russian sentences and represent the roles of participants and attributes of the certain event.

At the next step of our model formulated in the previous studies [3], the predicate system S is introduced. The system includes predicates $\mathrm{P}_i(x_i) \in S$, describing all possible values of the grammatical and semantic characteristics of the sentence words in a particular language.

The grammatical cases of nouns in the Russian language are specified via the predicate variable z:

$$\mathrm{P}(z) = z^{\mathrm{nom}} \vee z^{\mathrm{gen}} \vee z^{\mathrm{dat}} \vee z^{\mathrm{acc}} \vee z^{\mathrm{ins}} \vee z^{\mathrm{loc}}, \quad (2)$$

where nom, gen, dat, acc, ins, loc are nominative, genitive, dative, accusative, instrumental and prepositional cases, respectively.

We can also specify semantic features of the nouns, such as animality via the predicate variable x:

$$\mathrm{P}(x) = x^{\mathrm{anim}} \vee x^{\mathrm{inan}}, \quad (3)$$

where anim is animate, inan means an inanimate noun.

We specify the semantic categories, which can be recognized at NER step, via the predicate variable y:

$$\mathrm{P}(y) = y^{\mathrm{ORG}} \vee y^{\mathrm{PER}} \vee y^{\mathrm{LOC}} \vee y^{\mathrm{VEN}} \vee y^{\mathrm{TIME}} \vee y^{\mathrm{TOOL}} \vee y^{\mathrm{Others}}, \quad (4)$$

where ORG, PER, LOC, VEH denote organizations, person names, locations, and vehicles, respectively; TIME, TOOL denote date and/or time and tools used in an action, respectively; Others are used in case of impossible determination of the semantic attribute of a word.

To correctly select an Agent as the initiator of the action and an object to which the action is directed, we introduce three additional predicate variables *m, f,* and *l,*that formalize the passive voice in the Russian language.

$$\begin{aligned} \mathrm{P}(m) &= m^{\mathrm{Part}} \vee m^{\mathrm{NOT_{Part}}} \\ \mathrm{P}(f) &= f^{\mathrm{aux}} \vee f^{\mathrm{NOT_{aux}}} \\ \mathrm{P}(l) &= l^{\mathrm{suff}} \vee l^{\mathrm{NOT_{suff}}}. \end{aligned} \quad (5)$$

Multidimensional predicate $P(x, y, z, m, l, f)$ defines the roles of event arguments via the predicate variables, describing grammatical and semantic characteristics of words in sentences:

$$\begin{aligned} \mathrm{P}(x,y,z,m,l,f) \rightarrow\ & \mathrm{P}(x) \wedge \mathrm{P}(y) \wedge \mathrm{P}(z) \wedge \mathrm{P}(m) \\ & \wedge \mathrm{P}(l) \wedge \mathrm{P}(f) \\ \mathrm{P}(x,y,z,m,l,f) =\ & \gamma_k(x,y,z,m,l,f) \times \mathrm{P}(x) \times \mathrm{P}(y) \\ & \times \mathrm{P}(z) \times \mathrm{P}(m) \times \mathrm{P}(l) \times \mathrm{P}(f), \end{aligned} \quad (6)$$

where $k \in [1, h]$, $h = 6$ is the number of roles of event arguments considered in the model. They are Agent, Object, PLACE-ARG, TIME-ARG, INSTRUMENT-ARG, REASON-ARG. The predicate $\gamma_k(x, y, z, m, l, f) = 1$ if the specified characteristics of words in the phrase, which represents a certain event, define one of the above roles, and $\gamma_k(x, y, z, m, l, f) = 0$ if the conjunction of grammatical and semantic features of a word does not correspond to any of the roles. Then, relations between the characteristics of words in the sentence that do not describe any role of the event are excluded from the formula (6) by the predicate $\gamma_k(x, y, z, m, l, f)$.

We can specify the role of the Agent of the Event via the predicate γ_1. That is, the predicate γ_1 shows relations of grammatical and semantic characteristics of the words in Russian sentences that correspond to the Agent role of the CRIME, TRANSFER, and POLICE events:

$$\begin{aligned} &\gamma_1(x,z,m,l,f) \\ &= \left(y^{\mathrm{ORG}} \vee y^{\mathrm{PER}} \vee y^{\mathrm{VEN}} \vee y^{\mathrm{Other}}\right)\left(x^{\mathrm{anim}} \vee x^{\mathrm{inan}}\right) \\ &\wedge \left(z^{\mathrm{nom}}\left(f^{\mathrm{NOTaut}} l^{\mathrm{NOTsuff}} \vee m^{\mathrm{NOTPart}}\right)\right. \\ &\left. \vee z^{\mathrm{ins}}\left(f^{\mathrm{aux}} l^{\mathrm{suff}} \vee m^{\mathrm{Part}}\right)\right). \end{aligned} \quad (7)$$

The event Object is the second most core participant of the event after the Agent. Typically, in traditional grammar, it is a noun phrase that denotes the entity acted upon or which undergoes a change of state of motion. In our specific crime-related domain, the Object is more often a harmed person or a vehicle, which moves from one location to another, or something like this. We can also explicitly specify the role of the Event Object via the predicate γ_2:

$$\begin{aligned} &\gamma_2(x,y,z,m,l,f) \\ &= \left(y^{\mathrm{ORG}} \vee y^{\mathrm{PER}} \vee y^{\mathrm{VEN}} \vee y^{\mathrm{Other}}\right)\left(x^{\mathrm{anim}} \vee x^{\mathrm{inan}}\right) \\ &\wedge \left(z^{\mathrm{acc}} \vee z^{\mathrm{dat}}\right)\left(f^{\mathrm{NOTsuff}} l^{\mathrm{NOTsuff}} \vee m^{\mathrm{NOTPart}}\right) \\ &\vee z^{\mathrm{nom}}\left(f^{\mathrm{aux}} l^{\mathrm{suff}} \vee \mathrm{m}^{\mathrm{Part}}\right). \end{aligned} \quad (8)$$

In addition to the roles of participants of the event, we can identify other arguments via the LLESs. We distinguish the action attributes of PLACE-ARG and TIME-ARG via the predicates γ_3 and γ_4, respectively:

$$\gamma_3(x,y,z) = \left(y^{\mathrm{LOC}} \vee y^{\mathrm{Other}}\right) x^{\mathrm{inan}} z^{\mathrm{loc}} \quad (9)$$

$$\gamma_4(x,y,z) = y^{\mathrm{TIME}} \vee x^{\mathrm{inan}}\left(z^{\mathrm{loc}} \vee z^{\mathrm{acc}}\right). \quad (10)$$

In instances when there is not a word in a sentence, which satisfies these equations (7)-(10) we suppose that a TIME-ARG or a PLACE-ARG or even an object or an Agent is missing in this Event. For instance, in the sentence *"Yesterday, in the center of thetown, a deputy's carwas burned"*, the subject is missed.

C. CROSS-LINGUAL CREE TRANSFER TECHNIQUE

Cross-lingual CRE transfer technique is based on the fact that the same event may be described in various languages. If we have an event type and event arguments roles, which are covered in a sentence of the first part of the corpus, and additionally we have knowledge about the shared semantic space of aligned sentences in two languages, we can transfer

the type and arguments roles of the event from the source language sentence to the target language sentence.

At the first step of the target corpus part processing, we POS-tag raw texts employing morphological processing tools of a particular language. Next, drawing on the knowledge about CRE in a sentence of the source part of the corpus, we tag an event type and event arguments roles in an aligned sentence of the target corpus part. For this transferring, we found aligned sentences written in two languages in the two corpus parts and applied patterns of correspondence between morphological tags of the target language sentences and possible roles of the event participants and event attribute that we, in turn, can extract from the sentence of source corpus part.

The target language of the crime-related parallel corpus that we utilize in our experiment is the Kazakh language. This language is quite difficult for automatic processing. We suppose that the main reason for this is the agglutinativeness and highly inflectional of Turkic languages. This means that a single root may produce hundreds of word forms in the Kazakh language. Each word-forming morpheme has its own specific morphological or semantic meaning (for example, person, case, number, time, mood, etc.). Therefore, it is difficult, if possible, to create a training corpus of sufficient size with enough labeled events.

For that reason, annotating the Kazakh part of the corpus is based on the fact that the same event may be described in various languages, and labeled metadata of a sentence of one language can be transferred to an aligned sentence of another language. Thus, we have employed the knowledge about events and roles of the event arguments that are labeled in the Russian part of the parallel corpus to convey these labels to the Kazakh part of the corpus.

At the first step of the cross-lingual CRE transfer technique, we POS-tag Kazakh raw texts with morphological processing tools proposed by Makhambetov et al. [25]. Their method accounts for both inflectional and derivational morphology, including not fully productive derivation, and uses a standard HMM-based approach to disambiguate the Kazakh language.

As a result of morphological labeling, we have obtained tags with the complex morphological information that include both the POS-tag of the word root and the labels of morphological information represented by each morpheme. For example, in the <word pos = ''qyzmetker_R_ZE ler_N1i_S3nen_C6''> tag of ''qyzmetkerlerınen'' Kazakh word ''R_ZE'' label means common noun; ''N1'' means the morpheme of a plural noun; ''S3'' means a possessive case of the third person of a singular/plural noun and ''C6'' means an ablative case of a noun.

Next, taking as a basis the labeled texts of the Russian part of the corpus, we have labeled event types and roles of event participants and event attributes in the Kazakh part of the corpus.

In order to transfer labels from a source language sentence to the target language one, we create patterns of correspondence between morphological tags of the Kazakh sentence and possible triggers and roles of the event arguments, which we receive from the Russian sentence.

D. EVALUATION FRAMEWORK

The performance of the approach introduced in our paper has been ranked by traditional metrics. For each language of our parallel corpus, we calculated precision and recall individually. Considering the fact that we work with low-resource and under-annotated languages, and consequently, we do not have the corpora that include event annotation or corpora that can be used as the ''gold standard'', we are forced to employ experts to evaluate the results of our experiments. Nevertheless, our approach to computing recall and precision is based on the ACE (automatic content extraction) English Annotation Guidelines for Events [12] perspective on an event in general. In particular, the extent of an Event is always the entire sentence within a trigger of the Event occurs. Thus, calculating the recall of our experiment, we can assume that a sentence that comprises the trigger of a CRE describes this event.

Then following [12] arguments that an event can include only event participants or additional comprise attributes such as Time-ARG, PLACE-ARG, and INSTRUMENT-ARG, we consider two formats of the event. We tentatively call an event with includes a Predicate and participants of the event a ''*short event*'', and an event that comprises any of the event attributes in addition to the mentioned event element – as a ''*complete event*''.

Thus, to evaluate the correctness of our CREE approach, five hundred automatically extracted and identified CRE were randomly selected from each part of the corpus (source and target languages). To avoid potential bias and subjectivity, we involve two experts to analyze the source language and two experts for the target language. The experts were asked to confirm for each of the CRE the correctness of its type, trigger, Agent, Object, and event attributes. The rating scale allowed three values and can be described as follows: If the ''complete event'' was extracted correctly the expert marked it ''2''; If at least one of the attributes of the event was identified as incorrect but the trigger type and participants of the event were extracted correctly (''short event''), the expert marked it ''1''; otherwise the expert marked extracted CRE as ''0''.

Then, we calculated the precision of short and complete CRE extraction for source and target languages separately. To increase the validation of our study we calculated the agreement of experts via Cohen's kappa coefficients.

V. EXPERIMENT RESULTS

In order to successfully reproduce our experiment, along with an aligned parallel corpus and a dictionary that includes information about the lexis semantic class of the particular domain it is essential to produce the logical-linguistic equations that identify the respective roles of the event participants according to the predefined structure of the event subtype

TABLE 2. The distribution of event types in the source part of the corpus.

Event subtype	Original verb	Lemmatized verb	Stemmed verb
Injure	75	3984	3542
Offense	366	5178	3909
Movement	9	507	461
Traffic Accident	139	2351	2909
Arrest	239	9035	8221
Trial	231	4250	3804
PD	294	7433	6723

The triggers of events are verbs. There is a distribution of the original form, lemmatized, and stemmed verbs.

for particular target language of the corpus and the list of correspondence patterns between morphological tags of the target language and potential event arguments of the aligned sentence in the source language.

For our experiment, we leveraged spaCy, nltk, panas, nimpy and sklearn modules of Python. A demonstration of crime-related events extracted by our approach and Python programming codes is available at the GitLab repository.[2]

A. DATA COLLECTION

In our experiments as a crime-related parallel corpus, we utilize the bilingual corpus of two low-resource and under-annotated languages, namely Russian and Kazakh.

The corpus has been developed for more than three years [10], and now it includes row texts from four news websites on the Kazakhstan information Internet space zakon.kz, caravan.kz, lenta.kz, nur.kz for the period from April 2018 to June 2021. The choice of these bilingual sites stems from the fact that they provide a significant number of articles with criminal-related texts about various incidents, for example, robbery, murder, traffic accidents and others.

Now, the volume of the parallel Kazakh-Russian corpus is not very large and accounts for about 22,000 aligned sentences. In order to align sentences in the corpus, we have applied the automatic text alignment application based on the translation dictionary, followed by manual validation [10].

B. SOURCE PART OF THE CORPUS

Using verbs that occur both in the dictionary and in sentences as event triggers, we identify more than 30 thousand crime-related events in the source corpus part. As previously mentioned, an event type is determined according to the value of the element <DOMAIN> of the trigger verb in the dictionary. For example, following the dictionary, the verb 'stole' is a trigger for the Offense subtype of CRE.

Table 2 shows the distribution of these events into seven subtypes. We simultaneously consider the distribution of original verbs in sentences, verbs lemmatized and stemmed at the stage of preprocessing.

The imbalanced distribution of event types presented in Table 2 reflects the actual distribution of events in crime and police texts. Since we are using a pattern-based event extraction method (that is rule-based), the main problem with imbalanced data with the rationale of the results' generalizability to other domains. Thus, we position our approach and its results as focused on the crime and police news problem domain, which is also noted in the limitations of our study. Application of the proposed approach to other subject areas may require further adaptation. However, rule-based approaches are considered to be more flexible and adaptable as they can be updated or modified manually based on domain expertise or changes in data distribution, which of course can be time-consuming and require expert knowledge [50].

The triggers of events are verbs. There is a distribution of the original form, lemmatized, and stemmed verbs. The study of a verb as an event trigger confirms the obvious fact that the recall of the event extraction from the text is mostly higher in the case of considering the match of the dictionary form to the verb lemma in a text than in the analysis of the verb stems in the text. Taking into account the fact that a trigger can be not only a verb but also a noun, we considered about 500 nouns as triggers of events, determining the event type by the value of element <DOMAIN> of the noun in our dictionary.

Table 3 shows the distribution of events found in the source language part of the corpus into seven subtypes. We consider nouns, verbs, and noun + verb pairs as triggers separately. The utilization of a noun+verb pair as an event trigger can enhance precision and simultaneously decrease recall of CREE compared to using nouns and verbs as triggers of events separately. For example, using the two-word trigger 'court' + 'sentenced' can improve the precision of the TRIAL event subtype identification compared to using only the verb 'sentenced' as the event trigger. In total, about 4350 events are extracted from the Russian part of the corpus. Appendix B shows the sample of events extracted from the source language part of the corpus by applying verb+noun triggers.

TABLE 3. The distribution of events found in the source-language part of the corpus into seven subtypes.

Event type	Event subtype	Trigger type: noun lemma	verb lemma	noun + verb
CRIME	Injure	456	3984	298
	Offense	1972	5178	495
TRANSFER	Movement	132	507	69
	Traffic Accident	611	2351	104
POLICE	Arrest	947	9035	498
	Trial	1363	4250	1212
	PD	2217	7433	1653

In the next step, after selecting the events in the corpus and defining their types and subtypes, we identify the event arguments that include the participants and attributes of events described in 3.4. Using logical-linguistic equations (10)-(12), we determine Agent, Object, and the attribute roles in each selected action. Appendix C contains the sample of events extracted from the Russian part of the corpus, their subtypes, triggers, and arguments.

[2]https://gitlab.com/crime-event extraction

TABLE 4. The patterns of the Kazakh POS-tagging chunks that may correspond to the roles of the event arguments.

Roles of event arguments	POS-tags	Labels of cases	Label of possessive case
Agent	R_ZE, R_ZEQ, R_BOS	-	-
Object	R_ZE, R_ZEQ, R_BOS	C4, C2, C3	S*
PLACE-ARG	R_ZE, R_ZEQ	C5, C6, C3	-
TIME-ARG	R_ZE	C6	-
INSTRUMENT-ARG	R_ZE	C7, C3	S*

Trigger of Event	POS tags	Additional morphological information
Action	R_ET	Not ET_KSE and not ET_ESM and not ET_ETU and not ET_ETB

Here, R_ZE, R_ZEQ, R_BOS and R_ET are POS-tags of Noun, common; Noun, personal; Foreign word and Verb, accordingly. C2, C3, C4, C5, C6, C7 are cases of nouns. S* shows a possessive case.

C. TARGET LANGUAGE OF THE CORPUS

As the result of the POS-tag Kazakh language [25], we have derived tags with POS-tag of the word root and the labels of morphological information represented by each morpheme of the Kazakh language. To transfer labels from a source language sentence to the target language one, we create patterns of correspondence between morphological tags of the Kazakh sentence and potential event arguments of the aligned sentence in Russian.

Table 4 shows how morphological labels of the Kazakh text correspond to the potential roles of the event participants and event attributes. We base on the ''KazNLP: NLP tools for Kazakh language'' tagset.[3]

As a result of cross-lingual CRE transfer, we identified triggers, participants, and arguments of the events related to criminal or police work news from the Kazakh language part of our parallel corpus. Appendix D presents a sample of events, each of which includes triggers, subtypes, and arguments, that are extracted from the Kazakh part of the corpus.

In total, more than 450 events were extracted from the Kazakh part of the corpus, 69 events of them belong to the ARREST subtype, 72 CRE belong to the TRIAL subtype, 94 events to the INJURE subtype, 34 events to the TRAFFIC ACCIDENT subtype, 4 events of the MOVEMENT subtype, 102 CRE belong to the PD subtype, and 89 events to the OFFENSE.

D. EVALUATION OF THE RESULTS

Evaluation of the experiments results was carried out by two experts for each of the languages. Every one of them has more than ten years of experience in editorial or publishing activities. As we mentioned in section IV-D, the experts ranked five hundred randomly selected CREs that were automatically extracted from each part of the corpus (source and target languages) as ''0'', ''1'', ''or 2''. That enabled us to calculate precision for short and complete types of events by following a traditional equation:

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \tag{11}$$

here calculating the precision of extraction for the so-called ''complete'' events, we consider a true positive (tp) event that is marked as 2. And the false positive (fp) is the number of events that include attributes but were not marked by experts as ''2''.

According to our definition of a *complete event*, it includes event attributes additionally to event participants, in other words, a complete event comprises a short event plus event attributes. Bearing this in mind, for the so-called *short event*, we consider true positive (tp) as the number of events that are marked both ''2'' and ''1''. And the false positive (fp) is the number of events that were marked by experts as ''0''.

Table 5 shows the precision of CREE from the source and target languages of the parallel corpus, which are Russian and Kazakh, respectively.

To compute the recall of CREE, we based on our preceding assumption (Section IV) that if a sentence includes an event trigger, it should describe some CRE. We verify how many events were extracted from 500 randomly selected sentences with criminal-related triggers verbs from the corpus, and calculate recall by applying the following traditional equation separately for Kazakh and Russian languages:

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}. \tag{12}$$

In this case, we consider both the short and the complete event types as true positive (tp) CREE.

Table 6 presents the recall and F_1-measure of CREE from the parallel corpus for Russian and Kazakh languages, respectively.

Our analysis showed a decrease in the precision and obviously recall of target language processing compared with the source language. Most likely, the main reason for

TABLE 5. The precision of cree from parallel corpus.

Events	Source language (Russian)	Target language (Kazakh)
short CRE	76.30%	61.50%
complete CRE	73.00%	55.76%

TABLE 6. The recall and f_1-measure of Event Extraction from parallel corpus.

Measures/	Source language (Russian)	Target language (Kazakh)
recall CRE	94.80%	72.40%
F_1 short CRE	84.55%	66.51%
F_1 complete CRE	82.48%	63.00%

[3]https://github.com/nlacslab/kaznlp

TABLE 7. Comparison to the state of the art focusing on: various languages, the range of the Event Extraction types, and event arguments that were extracting.

Approaches and techniques	PR	RC	F_1	Languages	Types of CRE	Arguments of CRE
[8]	0.79	0.72	0.76	English	Cyber-security	Event subtypes
[50]	0.82	0.83	0.82	English	Hate crime	Only individual attributes
[24]	0.88	0.86	0.87	Malaysian	Drug-Related	Only individual attributes
[25]	0.64	0.68	-	Arabic	MANSA event	Short CRE
[9]	0.62	0.59	0.61	Indonesian	Various types	Complete CRE
PaCo-based approach	*0.76/0.73*	*0.94*	*0.85/0.82*	*Russian*	*Various types*	*Short CRE /complete CRE*
	0.62/0.56	*0.72*	*0.67/0.63*	*Kazakh*		*Short CRE/ complete CRE*

this is the agglutinativeness and polysemy of the morphology of the Kazakh language. Furthermore, the precision, and consequently F1-measure of ''short'' CRE extraction is higher than the ''complete'' CRE extraction, although only slightly.

Table 7 compares the effectiveness of proposed PaCo-based approach with other methods of CREE, focusing on various languages, the range of the event extraction types, and event arguments involved in the events.

Comparing the obtained recall, precision, and F_1 measure with preceding research [8], [34], we can approve that despite gaining not very high coefficients values, our results are comparable for the target language (Kazakh) and sometimes better for the source language (Russian).

However, in our case, we extract events that consist of all possible information about CRE (complete CRE), namely a type/subtype, trigger, Agent, Object, Time-ARG, PLACE-ARG, and INSTRUMENT-ARG.

Furthermore, unlike previous studies [8], [9], [25], [33], which considered only one specific type of event, for instance, hate crimes [33], and so on, we address a wide range of types and subtypes of CRE and calculate the extraction precision for all crime related event types together.

An additional advantage of our approach is an opportunity for event extraction from the texts in low-resource and under-annotated languages. To be able to refer to the results of experiments in the future, they must be as objective and accurate as possible. Generally, assessing the validity of experiments that are performed on a particular corpus is carried out either with the involvement of experts or by comparison with the so-called ''gold standard,'' i.e., the preliminarily annotated corpus.

Since the fact that our study concerns low-resource and under-annotated languages, we can't validate the results of experiments based on semantically pre-annotated corpora. For that reason, for research results evaluation the experts' opinions were used. Firstly, two native speaker experts independently assessed the short and complete CRE identification correctness and then the level of agreement of their opinions was checked using Cohen's kappa coefficient [49].

We calculated Cohen's kappa coefficients for two levels of events (short and complete CRE) in Russian and Kazakh languages separately. As previously mentioned, experts were asked to evaluate the results of the extraction, while the evaluation scale considered three possible options: 1 if short CRE was correctly identified, 2 if complete CRE was correctly identified and 0 if at least one of the event participants or the subtype of the event trigger was incorrectly identified. While it is worth noting that calculating the coefficient agreement of short CREs we considered correct events that were noted by the expert either 1 or 2.

TABLE 8. The confusion matrix of the experts' assessment: Extraction of short CRE.

	Source language (Russian)		Target language (Kazakh)	
First expert	Second expert			
	"0"	"1"	"0"	"1"
"0"	93	36	164	72
"1"	15	256	7	257

TABLE 9. The confusion matrix of the experts' assessment: Extraction of complete CRE.

	Source language (Russian)		Target language (Kazakh)	
First expert	Second expert			
	"0"	"2"	"0"	"2"
"0"	117	8	192	11
"2"	28	347	48	249

Tables 8-9 present the confusion matrix of the experts' assessment of the CREE from the Russian-Kazakh parallel corpus. In the tables, the rows present the decision of the first expert, and the columns, the decision of the second one for short and complete CREE of source and target parts of the parallel corpus.

Table 10 presents the Cohen's kappa coefficients for *short* and *complete* CRE of two languages separately that are calculated based on this matrix. The commonly accepted scale for estimating Cohen's kappa coefficient is as follows [49]: from 0.81 to 0.99 — near perfect agreement; from 0.6 to 0.80 — substantial agreement; from 0.41 to 0.60 — moderate agreement; and from 0.21 to 0.40 — fair agreement. Based on scale, we can claim that values of the Cohen's kappa coefficients contribute to increasing validation of our study. However, despite the fact that the values of the agreement coefficients look promising, obviously further study must continue to handle increasing semantically annotated corpora for validation of obtained results.

The summary of experiments evaluation measures is presented in the Figure 3.

TABLE 10. Cohen's kappa coefficients of agreement.

Events	Source language (Russian)	Target language (Kazakh)
short CRE	89.80%	84.20%
complete CRE	92.80%	88.20%

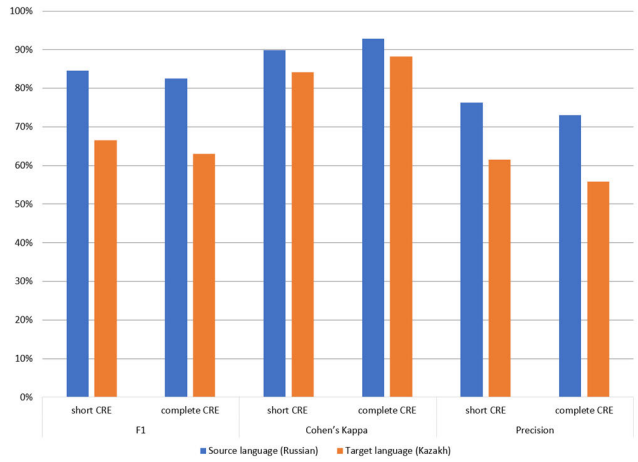


FIGURE 3. The summary of experiments evaluation measures.

VI. DISCUSSION

Although currently graph embedding methods are quite widely applied for event extraction from the text and sometimes these methods can be used even for low-resource languages, however, the limited training data and resources can pose a challenge for applying graph embedding methods effectively.

Moreover, for processing low-resource language usually transfer learning techniques are applied to adapt pre-trained graph embeddings from high-resource languages to low-resource languages. Transfer learning involves training a model on a source task (e.g., event extraction in a high-resource language) and then fine-tuning the model on a target task (e.g., event extraction in a low-resource language) with limited training data. However, this approach needs training data in high-resource language.

In our approach, we don't involve high-resource languages. In this study, we focused on the pattern-based EE approach that gives the opportunity to extract crime-related events from news articles that were published in low-resource and under-annotated languages.

First, the *major methodological contribution* of the work is the introduction of the two-stage method to extract criminal and police-related events from a bilingual parallel corpus, which is composed of two low-resource and under-annotated languages.

In the study we demonstrated how logical-linguistic equations, which represent roles of the event participants according to the predefined structure of the event subtype, and the Cross-lingual CRE transfer strategy could be successfully used for Crime-Related Event Extraction based on the parallel corpus.

As already noted, there are a large number of different techniques for event extraction. Most of them exploit pattern-based [14], [5] and machine learning methods [18], [20], [21]. The main reason for the absence of a unified standard approach for EE lies in the fact that ML approaches need large, semantically annotated corpora but a pattern-based event extraction approach is a time-consuming and labor-intensive task that must involve a lot of domains. Therefore, the *additional methodological contribution* of our research is the enhancement of pattern-based event extraction method [14], [16], [17], which is based on the multilingual synonyms dictionary with crime-related lexis and logic-linguistic equations. These equations allow us to represent the event's argument roles via the relationship between grammatical and semantic characteristics of the words in a sentence.

Regarding EE from the terrorism and criminal domain texts, on the one hand this domain can be considered a well-researched [3], [16], [17], but on the other hand, many of the involved studies consider the problem of CRE separately for various types of crime events (related to terrorism, cybercrime, crimes against the person, crimes related to transport, etc.) [8].

Enhancing the pattern-based event extraction method [8], [14], [17], we address the challenge of increasing the number of various event types related to police and criminal activities that can be extracted from news articles simultaneously.

As explained in Section IV, for modification of mentioned Cross-lingual technique, we propose to simultaneously use the (1) preliminary POS-tag labeling of target language texts; and (2) the patterns of the correspondence between POS-tags of target language sentences and possible roles of the event participants/attributes that are transferred from an aligned source language sentence. This modification allows us to handle the bilingual parallel corpus. The research [24] fairly states the EE task becomes more difficult for texts written in low-resourced and under-annotated languages.

Additionally, gold-standard annotations for event extraction are publicly available only for a few languages. Usually, in such corpora there are only in English [23] and some other European languages. In our study we *modify the cross-lingual CRE transfer technique* for processing the second part of the corpus (target language), based on supplementary knowledge about the semantic similarity patterns of the considered pair of languages [24].

The incremental *practical contribution* of the research is following. Unlike major studies on the detection and extraction of CRE in news articles, which analysed only one certain type of crime [33], [34], we consider the big group of events that relates to unlawful action. In order to detect these events, we have predefined the structure of three event types, namely, TRANSFER, CRIME, and POLICE several subtypes (see Table 1). Every subtype structure includes about two participants and several attributes of the action or event.

TABLE 11. Summary of research contributions.

Type of contribution	Contribution
Methodological and theoretical contributions	- Developing the two-stage method of extracting criminal and police-related events from a bilingual parallel corpus composed of two low-resource and under-annotated languages, we address the challenges of crime-related events extraction from not special domain documents (like news articles) described by [15, 50] - Enhancing the pattern-based event extraction method [21, 23, 24], we address the challenge of increasing the number of various event types related to police and criminal activities that can be extracted from news articles simultaneously. - Modifying the cross-lingual CRE transfer technique, we address the methodological challenges mentioned by language semantic similarity patterns researchers [34].
Practical contributions	- Predetermining the structure of seven subtypes of events allows extracting facts related to police and criminal activities from news websites clearer and more accuracy - Based on the modified cross-lingual CRE transfer technique, diverse EE applications for low-resourced and under-annotated languages can be designed. For example, we efficiently extracted CREs from news articles in the Kazakh language and obtained the patterns of POS-tags chunks of Kazakh texts that can represent event structures - The event subtypes distributions obtained as a result of the experiments can contribute to the development of the social research in regions (see Table III) - Using the special NLP approaches such as applying a pair of noun+verb as an event trigger instead of an only verb, as well as lemmatization of text verbs, allows increasing the facts extracting recall

Thus, one of the practical contributions is the distinguishing seven different subtypes of events that can be involved in a criminal action that allows obtaining facts related to police and criminal activities clearly and more accurately.

We should also highlight the practical contribution that was produced by the Cross-lingual CRE transfer technique for transferring labeled metadata from a sentence of one language into an aligned sentence of another language. Based on the technique, diverse EE applications for low-resourced and under-annotated languages can be designed. As explained in Section IV-C and summarized in Table 4, applying this technique for the Russian-Kazakh aligned corpus allows us to extract CREs from news articles in the Kazakh language. Every identified event comprises event type, roles of event participants and event attributes extracted from the Kazakh part of the corpus. Our experiment showed the precision of CREE in the Kazakh part of the corpus is 61.50% for short CRE that includes the correctly identified trigger, subtype/type, Agent, and Object, and 55.76% for complete CRE that includes extra correctly identified roles of the attributes of the event. We obtained the patterns of POS-tags chunks of Kazakh texts that can represent the event participants (Agent, Object) and event attributes (PLACE-ARG, TIME-ARG, and INSTRUMENT-ARG). Even though the obtained precision is lower than the average result of the Event Extraction approach [34], we have extracted CRE from texts in the Kazakh language for the first time. Since this language is a low-resource and under-annotated language, we had little capacity to involve extra-linguistic resources to process the Kazakh language.

Next, the event subtypes distributions obtained as a result of the experiments can contribute to the development of social research in regions. For instance, in our illustrative experiment on the dataset comprising texts on news articles of the Kazakh region (Table 3), CRE related to police activities appeared the most frequently (about 63%), and only about 12% and about 7% of events relate to directly suffered persons and traffic accidents, respectively. Including such handling for web news articles into various content analysis stages allows us to compare distributions of crime types of events by different countries and over different time periods.

Finally, it is worth mentioning the result of the research part concerning the problem of an event trigger identification in a sentence. Traditionally, the main verb of the sentence is considered as an event trigger in pattern-based event extraction approaches [2], [3]. However, based on the multilingual synonyms dictionary with criminal-related lexis (see Figure 2) and the conducted experiments, we demonstrate that the precision of CREE increases when a pair of a noun and a verb are considered as a trigger of the event. Additionally, as summarized in Table 2, we considered the impact of the verb form (original form, verbs lemmatized or stemmed verb) on the event extraction recall. We realize that we conducted experiments only on one text corpus. However, non-contradiction of observed results to the general NLP knowledge looks promising and allows us to expect our findings to be confirmed on other corpora.

In Table 11 we summarize the methodological and practical contributions of our research.

In the study, we experienced the following *limitations* (L) that will be addressed in future research correspondingly:

L1: the main limitation related to selecting the types and subtypes of events. Despite the fact that our approach allows extracting various event types, we considered only seven subtypes of events, namely Movement, Traffic Accident, Injure, Offense, Arrest, Trial, and Police Department. This choice was conditioned by the theme of our text corpus, which is generally focused on criminal topics.

L2: limitation related to the selection of the attributes of events. In the article, we considered only PLACE-ARG,

TABLE 12. Overview of approaches and techniques processed to Crime-Related texts

Approaches and techniques	Examples of the studies (references)	Methods	Dataset and processing language (if not English) or Regions	Effectiveness
Crime texts identification (Crime detection)	[11,12]	Various clustering techniques (Grid-based, constraint-based, k-means clustering algorithm, and others)	The specialized Communities and Crime dataset	F-measure reaches 87%
The crime event types classification	[35, 37]	Various ML techniques for classification (SVM and Neural Networks, and others)	Mostly, the specially annotated corpora: Annotated Crimes Corpus (Corpus Anotado de Delitos), Spanish corpus of Peruvian news English tweets dataset, Mexico	The average obtained F-Score result of classification lies between 77.9% and 84%
The crime event types classification	[35, 37]	Various ML techniques for classification (SVM and Neural Networks, and others)	Police- recorded crime event data, US Arrests dataset	Quite low. On average, precision reaches 0.50 with recall 0.16
Crime pattern modeling and crime prediction	[6, 30]	Various classification and clustering algorithms with additional time and spatial characteristics	Police crime reports and witness narrative reports for the USA, UAE, and India, a corpus of domestic violence events (New South Wales Police Force). The newspaper reports on crime against women in Indian states, spatial-temporally tagged tweets about crime events in Spanish language, and News reports from Malayalam online papers.	Detection of some type of CE (for instance, Weapons) for police crime or witness narrative reports achieves up PR - 0.96, RC - 0.90, and more. While F-measure was from 0.61 to 0.71 for CE identification in newspapers articles or social networks
Crime Information Extraction, including CE identification	[33, 34] [32]	Generally, rule-based language expression patterns combined with dictionary, ontology, and thesaurus were utilized Less often, cauterization methods were used (graph-based clustering technique)	Labelled corpora or often manually labelling data from Twitter, Instagram, Yahoo!, YouTube in English and Spanish, Dutch, Italian, Portuguese, Arabic, and some other languages	Accuracy achieved on average from 0.75 to 0.84, depending on a language. Some research showed that F-measure can achieve 0.9 on the good manually annotated tweets
Hate Speech Detection	[11, 13, 27, 29]	Supervised machine learning classifiers, Recurrent or Convolutional Neural Networks, BERT model, sometimes with exploitation lexical resources	Online newspaper articles from the USA and India; structured reports of events according to actor, city, and country level; manually labelled tweets; the information from Malaysian National News Agency (BERNAMA); manually annotated subsets of the local news articles	F-measure of CE extraction of various types of ranges from 0.64 (person) to 0.96 (drugs name). F-measure of extraction of complete crime events (covering trigger, type, and arguments) that relevant very narrow domain ranges from about 0.6 (cybersecurity events) to 0.64 (hate crime). And only when researchers used manually labelled corpus for training precision can be achieved about 0.83.

TIME-ARG, and INSTRUMENT-ARG arguments of events. REASON-ARG was not included within the scope of the study. We plan to consider the possibility of formalizing event reason attributes in future work.

L3: limitations related to elaborating the event annotation label set and their description. This section of the research is distinguished in a separate part of the study and will be considered in the near future.

L4: limitations related to demonstration of the research practical use. Here, we plan to implement and show in the next steps of the study the application of our methodology for solving particular problems of content analysis falling within occurrences of criminal and police activities.

L5: limitations related to possibility of results validating only via experts' opinions agreement coefficients. We assume that to solve the problem it is necessary the further development of the Russian-Kazakh parallel corpus and the other semantically annotated corpora.

VII. CONCLUSION

In this paper, we introduced the PaCo-based approach that allows extracting CRE (i) from the source language part of the parallel corpus by applying the *enhanced pattern-based method*; and (ii) from the target language of the corpus by applying the *cross-lingual CRE transfer technique* that is

based on supplementary knowledge about the semantic similarity patterns of the considered pair of corpus languages.

In this context, the main *methodological* contribution of the *EPB method* consists of: (1) the identification of the patterns of three main types and seven subtypes of events related to crime and police activities, namely ARREST, TRIAL, INJURE, TRAFFIC ACCIDENT, MOVEMENT, PD, and OFFENSE; the expansion of the crime-related and police-related type events list that can be extracted from news articles simultaneously (2) employing the logical-linguistic equations that determine the roles of the participants and attributes of events through the relations between the semantic and morphological characteristics of the sentence words in a particular language.

The main *methodological* contribution of the *cross-lingual CRE transfer technique* is provided by proof of the potential of utilizing the preliminary labeled events in one language texts to label events in another language texts of a parallel aligned corpus. Thus, our approach does not require a preliminarily annotated corpus for training making it more attractive to low-resource languages.

In our experiments, we utilized the bilingual crime-related aligned Russian-Kazakh corpus. As a main *practical* contribution of our research, the following two results can be determined. Firstly, based on extracting and labeling events of the parallel Russian-Kazakh corpus of news portals and social media articles, which are related to crime and police activities, we obtained the events distribution in the corpus according to the events subtypes. Secondly, we obtained the patterns of POS-tags chunks of Kazakh texts that can represent the event participants and event attributes; and also considered the possibility of using a verb, noun, and even a pair verb+noun as an event trigger.

In future work, the scientific contributions of our study can be applied to solve some problems of content analysis. For instance, to compare the distributions of CRE from different news sites in both the same language and various languages. Moreover, the distribution of these event types in various countries, as well as the distribution of them over time, can be considered.

APPENDIX A

See Table 12.

APPENDIX B

See Table 13.

APPENDIX C

See Table 14.

APPENDIX D

See Table 15.

TABLE 13. The fragment extracted events from the source part of the corpus (Screenshot)

File	sent number	Subtype of event	Trigger_noun	Trigger_verb
7670_ru_parsed	2	TRIAL	приговор	осуждены
7670_ru_parsed	2	TRIAL	УК	осуждены
1312_ru_parsed	13	TRIAL	УК	возбуждено
7887_ru_parsed	10	ARREST	показания	задержан
3708_ru_parsed	5	ARREST	задержанный	установлено
2596_ru_parsed	4	PD	полиция	обратились
1575_ru_parsed	9	INJURE	травма	умер
7854_ru_parsed	8	INJURE	ранение	нанесено
8704_ru_parsed	1	INJURE	убийство	подозревается
7991_ru_parsed	2	TRAFFIC_ACCIDENT	обгон	вылетел
9146_ru_parsed	11	TRAFFIC_ACCIDENT	наезд	совершил

TABLE 14. The example of events, each of which includes trigger, subtypes, and arguments, are extracted from the russian part of the corpus (Screenshot)

File	Sentence	Trigger of event	POS of trigger	Subtype of Event	Action	Agent	Object	PLACE-ARG	TIME-ARG	1st expert	2nd expert
8188_ru_parsed	0	убить	VERB	INJURE	убить	муж	женщина		с сентября	2	1
8704_ru_parsed	1	убить	VERB	INJURE	убить	сотрудник	жена	уральске		2	2
2555_ru_parsed	2	застрелить	VERB	INJURE	застрелить	полиция	подозреваемого	нападение		1	1
2018_ru_parsed	2	застрелить	VERB	INJURE	застрелить	оппонент			3 апреля 1998 года	2	0
2485_ru_parsed	2	сбить	VERB	INJURE	сбить	автомашина	ребенок	зебра		2	1
6507_ru_parsed	3	сбить	VERB	INJURE	сбить	машина	супруг		в 9:00 утра 12 ноября	2	2
3402_ru_parsed	0	грабить	VERB	INJURE	грабить	мужчина	несовершеннолетних	талдыкорган е		2	2

TABLE 15. The example of events, each of which includes triggers, subtypes, and arguments, are extracted from the kazakh part of the corpus (Screenshot)

File	Sentence	Event	Subtype of Event	Action	Agent	Object	PLACE-ARG	TIME-ARG	2nd expert
2728_kz_parsed	40	POLICE	TRIAL	танысуға	құқығыңыз	хаттамамен			1
2728_kz_parsed	40	POLICE	TRIAL	жазуға	құқығыңыз	хаттамамен			1
8542_kz_parsed	4	POLICE	PD	деп хабарлайды	қызметкерлері	қаласының			1
7995_kz_parsed	8	POLICE	PD	іздеуге кіріседі	полицейлер	хабарламаны			1
8188_kz_parsed	0	CRIME	INJURE	іздеуде болған	күйеуі	Қыркүйектен		әйелді	2
4703_kz_parsed	0	CRIME	INJURE	берді	Ауатуда				0
3977_kz_parsed	7	CRIME	OFFENSE	ішу		наурызда			0
3977_kz_parsed	7	CRIME	OFFENSE	жасаған		наурызда			1
8462_kz_parsed	4	CRIME	OFFENSE	көрсете		қағазын			1
8462_kz_parsed	4	CRIME	OFFENSE	алмаған		қағазын			1
5111_kz_parsed	4	POLICE	ARREST	тартып	Талдықорғанның				0
5111_kz_parsed	4	POLICE	ARREST	ұсталды	Талдықорғанның				1

REFERENCES

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

[2] J. Björne, F. Ginter, and T. Salakoski, "The biomedical event extraction downstream application," in *Proc. EPE*, 2017, pp. 17–24.

[3] J. A. Reyes-Ortiz, "Criminal event ontology population and enrichment using patterns recognition from text," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 11, Oct. 2019, Art. no. 1940014.

[4] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*, vol. 12. Berlin, Germany: Springer, 2012.

[5] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 9, no. 3, pp. 139–154, Jun. 2016.

[6] J. T. Nockleby, "Hate speech," in *Encyclopedia of the American Constitution*, 2nd ed. L. W. Levy and K. L. Karst, New York, NY, USA: Macmillan, 2000, pp. 1277–1279.

[7] A. H. Salas, J. Morzan-Samam, and M. Nunez-del-Prado, "Crime alert! crime typification in news based on text mining," in *Proc. Future Inf. Commun. Conf.*, in Lecture Notes in Networks and Systems, vol. 69, 2020, pp. 725–741.

[8] S. Yagcioglu, M. S. Seyfioglu, B. Citamak, B. Bardak, S. Guldamlasioglu, A. Yuksel, and E. I. Tatli, "Detecting cybersecurity events from noisy short text," 2019, *arXiv:1904.05054*.

[9] F. Rahma and A. Romadhony, "Rule-based crime information extraction on Indonesian digital news," in *Proc. Int. Conf. Data Sci. Its Appl. (ICoDSA)*, Oct. 2021, pp. 10–15.

[10] N. Khairova, A. Kolesnyk, O. Mamyrbayev, and K. Mukhsina, "The aligned Kazakh-Russian parallel corpus focused on the criminal theme," in *Proc. CEUR Workshop*, 2019, pp. 116–125.

[11] N. Khairova, A. Kolesnyk, and Y. Lytvynenko, "Automatic multilingual ontology generation based on texts focused on criminal topic," in *Proc. CEUR Workshop*, 2021, pp. 108–117.

[12] *Ace (Automatic Content Extraction) English Annotation Guidelines for Events*, Linguistic Data Consortium, Philadelphia, PA, USA, 2005.

[13] N. Khairova, S. Petrasova, and A. P. Gautam, "The logical-linguistic model of fact extraction from English texts," in *Information and Software Technologies* (Communications in Computer and Information Science), vol. 639. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-46254-7_51.

[14] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in *Proc. 11th Nat. Conf. Artif. Intell.*, 1993, pp. 811–816.

[15] F. Hogenboom, "Automated detection of financial events in news text," ERIM PhD Series in Research in Management, ERIM Reference Number: EPS-2014-326-LIS, SIKS Dissertation Series no. 2014-41.

[16] Q. Li, Q. Zhang, J. Yao, and Y. Zhang, "Event extraction for criminal legal text," in *Proc. IEEE Int. Conf. Knowl. Graph (ICKG)*, Aug. 2020, pp. 573–580.

[17] F. Abdelkoui and M.-K. Kholladi, "Extracting criminal-related events from Arabic tweets: A spatio-temporal approach," *J. Inf. Technol. Res.*, vol. 10, no. 3, pp. 34–47, Jul. 2017.

[18] L. Sha, J. Liu, C.-Y. Lin, S. Li, B. Chang, and Z. Sui, "RBPB: Regularization-based pattern balancing method for event extraction," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 1224–1234.

[19] W. Xiang and B. Wang, "A survey of event extraction from text," *IEEE Access*, vol. 7, pp. 173111–173137, 2019.

[20] C. D. Manning, "Computational linguistics and deep learning," *Comput. Linguistics*, vol. 41, no. 4, pp. 701–707, Dec. 2015.

[21] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1247–1256.

[22] J. Liu, Y. Chen, K. Liu, and J. Zhao, "Event detection via gated multilingual attention mechanism," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4865–4872.

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

[24] S. Fincke, S. Agarwal, S. Miller, and E. Boschee, "Language model priming for cross-lingual event extraction," 2021, *arXiv:2109.12383*.

[25] K. T. Hossain, S. Gao, B. Kennedy, A. Galstyan, and P. Natarajan, "Forecasting violent events in the middle east and north Africa using the hidden Markov model and regularized autoregressive models," *J. Defense Model. Simulation, Appl., Methodology, Technol.*, vol. 17, no. 3, pp. 269–283, Jul. 2020.

[26] P. Chen and J. Kurland, "Time, place, and modus operandi: A simple apriori algorithm experiment for crime pattern detection," in *Proc. 9th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Zakynthos, Greece, Jul. 2018, pp. 1–3.

[27] M. Siino, D. N. Elisa, I. Tinnirello, and M. L. Cascia, "Detection of hate speech spreaders using convolutional neural networks," in *Proc. CLEF Labs Workshops*, 2021, pp. 2126–2136.

[28] K. Miok, B. Skrlj, D. Zaharie, and M. Robnik-Sikonja, "To BAN or not to BAN: Bayesian attention networks for reliable hate speech detection," *Cognit. Comput.*, vol. 14, pp. 1–19, Jan. 2021.

[29] K. A. Qureshi and M. Sabih, "Un-compromised credibility: Social media based multi-class hate speech classification for text," *IEEE Access*, vol. 9, pp. 109465–109477, 2021.

[30] P. Das and A. K. Das, "Graph-based clustering of extracted paraphrases for labelling crime reports," *Knowl.-Based Syst.*, vol. 179, pp. 55–76, Sep. 2019.

[31] T. Dasgupta, A. Naskar, R. Saha, and L. Dey, "CrimeProfiler: Crime information extraction and visualization from news media," in *Proc. Int. Conf. Web Intell.*, Aug. 2017, pp. 541–549.

[32] P. Das and A. K. Das, "A two-stage approach of named-entity recognition for crime analysis," in *Proc. 8th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2017, pp. 1–5, doi: 10.1109/ICCCNT.2017.8203949.

[33] K. R. Rahem and N. Omar, "Drug-related crime information extraction and analysis," in *Proc. 6th Int. Conf. Inf. Technol. Multimedia*, Nov. 2014, pp. 250–254.

[34] A. Mostafazadeh Davani, L. Yeh, M. Atari, B. Kennedy, G. Portillo Wightman, E. Gonzalez, N. Delong, R. Bhatia, A. Mirinjian, X. Ren, and M. Dehghani, "Reporting the unreported: Event extraction for analyzing the local representation of hate crimes," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5753–5757.

[35] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: A review," *IEEE Access*, vol. 9, pp. 88364–88376, 2021.

[36] *Rich ERE Annotation Guidelines Overview*, Linguistic Data Consortium, Philadelphia, PA, USA, Aug. 2021. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016T23

[37] A. Ramponi, B. Plank, and R. Lombardo, "Cross-domain evaluation of edge detection for biomedical event extraction," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1982–1989.

[38] F. F. Rosa, M. Jino, and R. Bonacin, "Towards an ontology of security assessment: A Core model proposal," in *Proc. 15th Int. Conf. Inf. Technol. Inf. Technol.-New Gener.*, vol. 738. Springer, 2018, pp. 75–80.

[39] Ç. Çöltekin, "A corpus of Turkish offensive language on social media," in *Proc. 12th Conf. Lang. Resour. Eval. (LREC)*, 2020, pp. 6174–6184.

[40] R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, "Aggression-annotated corpus of hindi-English code-mixed data," 2018, *arXiv:1803.09402*.

[41] D. Battistelli, C. Bruneau, and V. Dragos, "Building a formal model for hate detection in French corpora," in *Proc. 24th Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, vol. 176, 2020, pp. 2358–2365.

[42] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, "SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020)," in *Proc. 14th Workshop Semantic Eval.*, 2020, pp. 1425–1447.

[43] S. Goźdź-Roszkowski, "Corpus linguistics in legal discourse," *Int. J. Semiotics Law Revue Internationale de Sémiotique Juridique*, vol. 34, no. 5, pp. 1515–1540, Nov. 2021, doi: 10.1007/s11196-021-09860-8.

[44] S. Mukherjee and K. Sarkar, "Analyzing large news corpus using text mining techniques for recognizing high crime prone areas," in *Proc. IEEE Calcutta Conf. (CALCON)*, Feb. 2020, pp. 444–450, doi: 10.1109/CALCON49167.2020.9106554.

[45] A. Adily, G. Karystianis, and T. Butler, "Text mining police narratives for mentions of mental disorders in family and domestic violence events," in *Trends and Issues in Crime and Criminal Justice*, no. 629. Canberra, ACT, Australia: Australian Institute of Criminology, 2021, doi: 10.52922/ti04930.

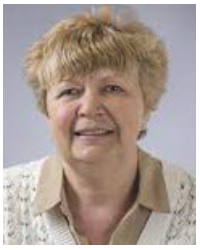
[46] G. Karystianis, A. Adily, P. Schofield, L. Knight, C. Galdon, D. Greenberg, L. Jorm, G. Nenadic, and T. Butler, "Automatic extraction of mental health disorders from domestic violence police narratives: Text mining study," *J. Med. Internet Res.*, vol. 20, no. 9, Sep. 2018, Art. no. e11548, doi: 10.2196/11548.

[47] R. Resende de Mendonça, D. Felix de Brito, F. de Franco Rosa, J. C. dos Reis, and R. Bonacin, "A framework for detecting intentions of criminal acts in social media: A case study on Twitter," *Information*, vol. 11, no. 3, p. 154, Mar. 2020, doi: 10.3390/info11030154.

[48] H. Chen, J. Schroeder, R. V. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, C. Boarman, K. Rasmussen, and A. W. Clements, "COPLINK connect: Information and knowledge management for law enforcement," *Decis. Support Syst.*, vol. 34, no. 3, pp. 271–285, Feb. 2003.

[49] A. S. Kolesnyk and N. F. Khairova, "Justification for the use of Cohen's Kappa statistic in experimental studies of NLP and text mining," *Cybern. Syst. Anal.*, vol. 58, no. 2, pp. 280–288, Mar. 2022.

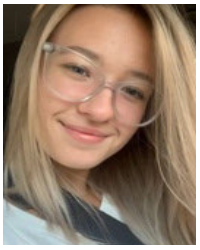
[50] P. P. Angelov, *Evolving Rule-Based Models: A Tool for Design of Flexible Adaptive Systems*, vol. 92. Berlin, Germany: Springer, 2002.



NINA KHAIROVA received the Ph.D. degree in information technology from the Kharkiv National University of Radio Electronics, Ukraine. She carried out her postdoctoral studies in computational linguistics with the National Technical University ''Kharkiv Polytechnic Institute'' (NTU ''KhPI''), Ukraine, and Taras Shevchenko National University, Ukraine. Since 2013, she has been a Professor with the Intelligent Computer Systems Department, NTU ''KhPI.'' She is currently a Visiting Professor with the Computer Science Department, Umeå University, Sweden. Her main research interests include corpus linguistics, natural language processing, machine learning, and artificial intelligence. She is the author of three books and more than 150 publications.

ORKEN MAMYRBAYEV received the B.S. and M.S. degrees in information systems from Abai University, Almaty, Kazakhstan, and the Ph.D. degree in information systems from Kazakh National Technical University named after K. I. Satbayev. He was an Associate Professor with the Institute of Information and Computational Technologies, Kazakhstan. He has been a Senior Researcher with the Laboratory of Computer Engineering of Intelligent Systems, Institute of Information and Computational Technologies. He is currently the Deputy General Director and the Head of the Laboratory of Computer Engineering of Intelligent Systems, Institute of Information, Kazakhstan. He is also a member of the Dissertation Council ''Information Systems,'' L. N. Gumilyov Eurasian National University in the specialties computer sciences and information systems. He is the author of five books, more than 130 articles, and more than 20 inventions and copyright certificates for an intellectual property object in software. His main research interests include machine learning, deep learning, and speech technologies.

NINA RIZUN received the Ph.D. degree in technical sciences from the Dnipro University of Technology, Ukraine. She is currently an Assistant Professor with the Department of Informatics in Management, Faculty of Management and Economics, Gdańsk University of Technology, Poland. Within the last years, she has also focused on the field of investigating ICT-related governance domains, including smart governance, smart cities, and open data. She has a strong ongoing publication record as a senior and second author in high-ranking international peer-reviewed journals such as IEEE, *Computers in Industry*, *Business Process Management Journal*, and *Telecommunication Policy*. Her main research interests include computational linguistics, NLP, sentiment analysis, artificial intelligence, and machine learning and their application for business processes and service improvement in such domains as healthcare and IT service management. She is an Editorial Board Member of the *eJournal of eDemocracy and Open Government* (JeDEM). She is a reviewer of several peer-reviewed journals. She is a program committee member of many high-level international conferences.



MARIIA RAZNO received the B.S. and M.S. degrees in applied linguistics from the National Technical University ''Kharkiv Polytechnic Institute'' (NTU ''KhPI''). She is currently pursuing the Ph.D. degree with the Slavistics Department, Jena National University. After graduation, she was an Assistant Professor with the Intelligent Computer Systems Department, NTU ''KhPI.'' She completed Grammarly Summer School, Kyiv, Ukraine, and specially completed natural language processing course with the Kharkiv National University of Radio Electronics. She has a lot of experience working with named entity recognition models and text classification models. She is the coauthor of four publications on discourse analysis and text classification topics. Her research interests include discourse analysis, text generation, speech to text, and text to speech models.

YBYTAYEVA GALIYA received the B.S. degree in information systems from Kazakh-British Technical University, Almaty, Kazakhstan, and the M.S. degree in information systems from Kazakh National Research Technical University named after K. I. Satpayev, Almaty, where she is currently pursuing the Ph.D. degree in management information systems.

She is a Researcher with the Laboratory of Computer Engineering of Intelligent Systems, Institute of Information and Computational Technologies, National Academy of Sciences of the RK, Almaty. Her research interests include corpus linguistics, natural language processing, machine learning, and speech technologies.