Event-Based Knowledge MLM for Arabic Event Detection

Asma Yamani[2], **Amjad Alsulami**[2] and **Rabeah Al-Zaidy**[1, 2]
[1]Center for Integrative Petroleum Research (CIPR)
[2] Information and Computer Science Department
King Fahd University of Petroleum and Minerals
Saudi Arabia
{g201906630,g202101130,rabeah.alzaidy}@kfupm.edu.sa

Abstract

With the fast pace of reporting around the globe from various sources, event extraction has increasingly become an important task in NLP. The use of pre-trained language models (PTMs) has become popular to provide contextual representation for downstream tasks. This work aims to pre-train language models that enhance event extraction accuracy. To this end, we propose an Event-Based Knowledge (EBK) masking approach to mask the most significant terms in the event detection task. These significant terms are based on an external knowledge source that is curated for the purpose of event detection for the Arabic language. The proposed approach improves the classification accuracy of all the 9 event types. The experimental results demonstrate the effectiveness of the proposed masking approach and encourage further exploration.

1 Introduction

Our lives are a sequence of events. Some of them concern the individual, some have their effect extended to a greater population, where others can even have a global effect. As the sources of news about events vary and the speed of the reporting has increased dramatically, event extraction has become an important challenge for governments and different agencies to have appropriate responses to the concerning events. Event extraction composes mainly of 2 tasks. The first is *event detection*, in which the event is detected, usually by a trigger, and then classified. A subsequent task is *event argument extraction*. It aims to identify different semantic entities related to the detected and classified event. There are several challenges related to event extraction and annotation, such as having multiple event types for the same piece of news, i.e. *Multi-label* problem. Also, multiple roles for the same entity, commonly referred to as the *role overlap* problem. In addition, similar sentences that contain the event trigger and the same entities may be classified as being an event or not based on the *tense*, whether it is an event that happened or something that is planned for in the future. All these challenges contribute to the complexity of the event extraction problem.

As with many downstream tasks, a sophisticated text representation, through contextual representation and attention mechanisms, was able to improve the performance of event detection models as shown in various studies related to events in the English language (Yang et al., 2019; Wang et al., 2019; Caselli et al., 2021). However, this has not been widely explored yet in event extraction for events reported in the Arabic Language. Event detection studies related to Arabic Language mainly focus on feature extraction using statistical approaches such as TF-IDF (Chouigui et al., 2018) and N-gram along with Part-of-Speech (POS) and Named Entity Recognition (NER) (Smadi and Qawasmeh, 2018; Alsaedi and Burnap, 2015) or using rule-based approaches as in (Mohammad and Qawasmeh, 2016).

In addition, domain adaptation through continuing to pre-train a contextual model, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), on domain-specific corpora such as events (Caselli et al., 2021), or modifying the masked language modeling (MLM) learning task to focus on entities have shown significant effect on the performance of the downstream task that they are catered towards (e.g., NER, medical domain NLP tasks, stance detection tasks) (Lin et al., 2021; Kawintiranon and Singh, 2021). As imposing an inductive bias to the MLM learning task has yet to be explored for the event extraction task, or when modeling the Arabic Language, we propose Event-Based Knowledge Masked Language Model (EBK-MLM) for the purpose of better detection and classification of events reported in the Arabic Language.

The contribution of this work is as follows: (1) We collect and annotate an Arabic Event dataset

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 273 - 286
December 8, 2022 ©2022 Association for Computational Linguistics

namely AraEvent[1] which consists of 2 sub-datasets AraEvent(November) and AraEvent(July) that is sourced from 4 popular Twitter news accounts. (2) We customize the MLM learning task to have an inductive bias towards the most significant terms in events, which achieves an average of 3.67% accuracy improvement in the event detection and classification task when tested on a non-homogeneous dataset, with up to 6.25% improvement in some event types[2].

The rest of this paper is structured as follows: in Section 2, we present some related work to our study. Then in Section 3, we propose the pre-training method. In Section 4, we present our fine-tuning experiments, whose results we discuss in Section 5. In section 6, we list the limitations of our work. Finally, in Section 7, we conclude and set some future directions.

2 Related Work

2.1 Event Extraction

Event detection, the first component of event extraction, usually starts by identifying the trigger, which is the word that most clearly identifies an event type, then the event classification task would follow (Chen et al., 2015; Nguyen and Grishman, 2015; Liu et al., 2016; Chen et al., 2017). However, more recent work (Liu et al., 2019) focuses on detecting the event without identifying the trigger as some events do not contain triggers. In addition, annotating the clearest trigger is a time-consuming task. The study proposed Type-aware Bias Neural Network with Attention Mechanisms that takes as an input a tokenized sentence with NER tags coupled with the event type then builds the representation based on the target event type. The output is 1 if the sentence conveys the event type, $zero$ otherwise. The attention mechanism gave more weight to the trigger words when developing the representation. The resulting model had similar performance on the ACE2005 event extraction dataset to SOTA event detection models (Sun et al., 2019; Chen et al., 2015; Nguyen and Grishman, 2015; Liu et al., 2016; Chen et al., 2017) that started with identifying the trigger, however, without using attention.

Using the representation of pre-trained contextual language models such as BERT for different downstream tasks, and more specifically here the event extraction task, have been gaining attraction recently. In (Yang et al., 2019; Wang et al., 2019) fine-tuned BERT is used for event argument extraction. The first study (Yang et al., 2019) identifies the trigger first via multiple fine-tuned BERT models for sentence classification, then based on the class(es) of events triggered, the arguments are extracted via a second BERT component fine-tuned for token classification to extract the arguments. In (Wang et al., 2019), a hierarchical approach is applied, in which the instance embedding from the BERT module for each token is concatenated with a rule-oriented embedding generated by hierarchical modular attention to classify Person, Time, Organization and Location. The result from this classification is finally fed to the Argument Role Classifier. The study (Caselli et al., 2021) follows a domain adaptive retraining approach, in which it continues pre-training BERT from the *'bert-base-uncased'* checkpoint on $79,515$ articles containing news about past or ongoing protest-related events. This improves the Trigger detection $F1$ score from 0.41, when using $BERT$, to 0.73 when using the $PROTEST-ER$ model that is pre-trained on protest-related articles. It also improves the argument extraction $F1$ score from 0.20, when using $BERT$, to 0.42 when using the $PROTEST-ER$ model. Our work aims to adapt the MLM task to give higher significance to words related to the events of our interest.

2.2 Arabic Event Extraction

In recent years, event detection and extraction systems that support the Arabic Language have evolved gradually. In a study, an event detection framework is introduced, which aims to detect disruptive events using temporal, textual, and spatial features (Alsaedi and Burnap, 2015). First, to differentiate between event and non-events tweets, a Naive Bayes classifier is trained and tested on a dataset that consists of 1200 tweets. The words composing the tweet are taken into account as features with the attributes: Unigrams, Bigrams, POS, NER. Compared to SVM and Logistic Regression, Naive Bayes performed the best, achieving an $F1$ score of 0.80. An unsupervised rules-based approach is proposed to extract events from Arabic tweets (Mohammad and Qawasmeh, 2016). Extracting the event, demystifying the NER and the Temporal resolution are all the three stages mentioned to extract the event. Focusing on event detec-

[1] https://huggingface.co/datasets/Asma/AraEvent

[2] https://huggingface.co/Asma/EBK-BERT

tion phase, Automatic Content Extraction (ACE) guidelines are mapped into syntax rules that use POS tags to extract event statements, event triggers, event time, and event type. For evaluation, $1,000$ Arabic tweets are used to evaluate the proposed approach, which maintained a 75.9% accuracy for extracting event triggers using Naive Bayes. Another study (Smadi and Qawasmeh, 2018) extracts a set of features from tweets for the events extraction task. Morphological features are used to analyze the structure of the text. POS, semantic features like NER, and word features such as Unigrams and their TF-IDF represents the different sets of features extracted by the system. To evaluate the proposed approach, a dataset of 2k Arabic tweets is utilized, and three classifiers are used: SVM, Naive Bayes, and Decision Tree. Results shows SVM scoring the highest F1 score for the event trigger extraction task scores with 92.6%.The study (Chouigui et al., 2018) presents statistical approaches for the event extraction task.

Focusing on Arabic news articles' titles, keywords are extracted by calculating the term weight for each word utilizing TF-IDF and comparing it with a threshold. For each keyword extracted, the event is defined using the POS co-occurrence rule. To evaluate the system, another news site is used for the events extraction task. The results shows that the performance of the approach is class-based and works well for domain-specific events such as the economy. As for datasets, EveTAR (Almerekhi et al., 2016) is the first publicly-available Arabic event detection dataset. In total, there are 590M tweets covering 66 significant events (eight categories). Using Wikipedia's Current Events Portal, it was collected over a one-month period. Tweets related to an event are grouped according to their time period of occurrence in order to represent that event. After cleaning and removing inaccessible tweets belonging to inaccessible accounts, the second version of the dataset comprised of 355M tweets (Hasanain et al., 2017). A recent study (Alharbi and Lee, 2021) presents a multi-dialect Arabic Twitter corpus for crisis events that include more than a million Arabic tweets from 22 crises and hazards between 2018 and 2020. To benchmark the dataset, AraBERT base model is fine-tuned by using annotated data from the same event to categorize tweets according to different labels. Despite limited task-specific training data, BERT-based models perform well on this task. Transformer-based models have yet to be used or evaluated for the detection and extraction of Arabic events of various types.

2.3 Arabic Pre-Trained Language Models (PTMs)

Pre-trained contextual representation models are known to be well suited for tasks that require understanding a given text, such as sentiment analysis, NER, and extractive question answering. One of the first Arabic PTMs with BERT$_{base}$ architecture is AraBERT(Antoun et al., 2020). It uses the BERT$_{base}$ configuration (Devlin et al., 2018) and is trained on both the MLM and Next sentence prediction tasks (NSP). Other Arabic PTMs trained on the BERT configurations are QARiB(Abdelali et al., 2021), MARBERT, ARBERT(Abdul-Mageed et al., 2021), and CAMeLBERT(Inoue et al., 2021). They mainly differ in the pre-training data source, such as whether they included Dialectal Arabic (DA), e.g., QARiB, MARBERT, CAMeLBERT-DA, CAMeLBERT-Mix or used only Modern Standard Arabic (MSA), e.g., AraBERT, ARBERT. Other differences include the size of the pre-training data and the ratio of DA to MSA, e.g. CAMeLBERT and QARiB. However, changing the masking procedure for the MLM training task has not been, to the best of our knowledge, investigated for Arabic Language.

Also, although multiple studies utilize PTMs in various tasks and applications, the use of contextual representation for event extraction in Arabic, has not been investigated yet.

2.4 Variations in the MLM learning task for PTMs

Masked Language Modeling (MLM) is a training task in which a model tries to learn the masked token representation using the surrounding unmasked words. It is adopted by BERT (Devlin et al., 2018) to train Language models by masking 15% of the tokens in pre-training. In BERT, 80% of the masked tokens are masked by $[Mask]$, 10% by the original token, and 10% by a random token (Devlin et al., 2018). Several studies have varied the masked token selection for the MLM training objective. A study proposed BERTSpan that masks contiguous random spans (Joshi et al., 2019). BERTSpan outperforms BERT on the extractive question answering tasks, coreference resolution, and 9 GLUE tasks. In (Sun et al., 2019), Enhanced Representation through kNowledge IntEgration (ERINE) is proposed to mask phrases and entities rather than random tokens. ERNIE is applied to 5 Chinese NLP tasks, including natural language inference, semantic similarity, named entity recognition, sentiment analysis, and question answering and it improved NER and natural Language inference the most.

In Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis (SKEP)(Tian et al., 2020), Pointwise Mutual Information is used to identify the most important words for the sentiment analysis task. It outperforms RoBERTa on sentence-level sentiment classification, Aspect-level sentiment classification, and opinion role labeling. Knowledge Enhanced Masked Language Model (KE-MLM) is proposed for Stance Detection (Kawintiranon and Singh, 2021). It uses log-odds-ratio to identify key stance tokens then used them for selecting the mask. It outperforms SKEP and other BERT variations on the Stance Detection task. Another specific application for knowledge-based masking is in the medical domain, in which Medical Entities are masked (Lin et al., 2021). It outperforms random masking and other baseline models on three clinical NLP tasks, TLINK temporal relation extraction, DocTimeRel classification, and negation detection, and one biomedical task, PubMedQA. Using cloze-like masking is proposed to provide indirect supervision to downstream tasks in a self supervised setting (Zhang and Hashimoto, 2021). The approach is evaluated on three text-classification tasks by masking words that exhibit a strong indication for the classes of the downstream task during a second stage pre-training of BERT. The results showed improved performance of models using cloze-like masking over other contextual models not masked using cloze-like masking. As knowledge-based masking is not addressed for the purpose of event extraction, in this work, we aim to leverage the knowledge related to certain types of events in the masking process in order to improve the representation of word related to our down-stream task.

3 Pre-training EBK-BERT

We propose Event Knowledge-Based BERT (EBK-BERT), which leverages knowledge extracted from events-related sentences to mask words that are significant to the events detection task (Section 3.1). This approach aims to produce a language model that enhances the performance of the down-stream event detection task, which is later trained during the fine-tuning process. The BERT-base configuration is adopted which has 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length, and a total of $110M$ parameters. The details of the implementation is in the following subsections.

3.1 EBK Token Masking

As previous studies have shown, contextual representation models that are pre-trained using the MLM training task benefit from masking the most significant words, using whole word masking. To select the most significant words we use odds-ratio (Szumilas, 2010). Only words with greater than 2 odds-ratio are considered in the masking, which means the words included are at least twice as likely to appear in one event type than the other. Calculating the odds-ratio for event detection is calculated as:

$$logodds(w, e) = \frac{\|e \text{ and } w\| \times \|!e \text{ and } !w\|}{\|!e \text{ and } w\| \times \|e \text{ and } !w\|} \quad (1)$$

were w is the word we are calculating the log-odds ratio for, with respect to a particular event e. Top 5 significant words are presented in Table 1

In order to mitigate the effect of noise generated by rare words, we perform word lemmatization using the Farasa lemmatizer (Abdelali et al., 2016), which combines, to a great extent, different word surfaces to their lemma. As presented in Appendix A Table 9, the vocabulary size shrinks after lemmatization. It combines words such as الفيضان, فيضانات, and فيضان, into one word فيضان, which helps focus the mask later on the most significant part of the word and avoid inflated odds-ratio values due to the infrequent terms. It is worth noting that there are words that appear in 2 or, at maximum, 3 event types. Event types *Contact*, followed by *Personnel* and *Nature* most significant words have the highest presence in the pre-training corpus based on 8 million sentences drawn randomly. The density of the frequency of the words is: 78.7% of the words are composed of one token, 19.7% of the words are composed of two tokens, and less than %2 words compose of more than 2 tokens.

3.2 Pre-training Data

The pre-training data consists of news articles from the 1.5 billion words corpus by (El-Khair, 2016). Due to computation limitations, we only use articles from Alittihad, Riyadh, Almasryalyoum, and

top	Personnel	Transaction	Contact	Nature	Movement	Life	Justice	Conflict	business
1	استقال	راجحي	التقى	ارضي	اجلاء	مقتل	قبض	اشتباك	انشاء
2	اقال	توصيل	نظير	فيضان	غادر	قتيل	اتهم	معركة	نسخ
3	رشح	استحواذ	لقاء	درجة	مغادر	اصاب	أطاح	قصف	اختار
4	ترشح	تمويل	خادم	درج	هجر	حالة	ضبط	معرك	شراء
5	مهمة	نيوكاسل	بحث	اعصار	نزوح	تسجيل	ايقاف	نفاية	تسلا

Table 1: Top 5 significant words (after Farasa's lemmatization) using odds-ratio

Alqabas, which amount to $10GB$ of text and about 8M sentences after splitting the articles to approximately 100 word sentences to accommodate the 128 max_sentence length used when training the model. The average number of tokens per sentence is 105. The normalization is performed as described in Section 4.1.1.

3.3 Preparing Data for BERT Pre-training

A WordPiece (Schuster and Nakajima, 2012) tokenizer is trained on the entire dataset ($10GB$ text) with a vocabulary size of 30522 using Hugging Face's tokenizers. For the baseline model, 15% of the tokens, are randomly masked with $[MASK]$. For the EBK-BERT model, 10% of the tokens are masked randomly and the remaining 5% are masked by considering the top $80 - 100$ words from each event type ordered by the odds-ratio.

3.4 Pre-training Setup

Google Cloud GPU is used for pre-training the model. The selected hyperparameters are: learning rate=$1e-4$, batch size =16, maximum sequence length = 128 and average sequence length = 104. In total, we pre-trained our models for $500,000$ steps, completing 1 epoch. Pre-training a single model took approximately 2.25 days.

3.5 Pre-training Results

Due to computation limitations, the model is trained for 1 epoch. We notice from Figure 1 that *EBK-BERT* has lower training loss than the *RandMask* model. This, however, cannot be an indicator to the performance of the model as $1/3$ of the masked words, which the model is learning the representation for, focus on about 3000 words from the 9 event types, whilst for the *RandMask* all the 15% masked words are random which adds complexity to the training process.

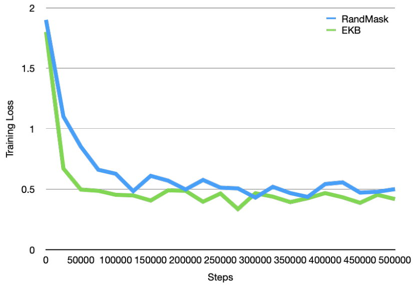


Figure 1: Training loss of *EBK-BERT* and *RandMask* model.

4 Fine-tuning Experiment

4.1 Event Data

The events dataset construction process is comprised of three steps: (1) Scrape tweets from four well-known Arabic news accounts on Twitter. (2) Conduct cleaning and filtering procedures on the collected tweets by applying text normalization. (3) Perform the annotation task by labeling the tweets according to their content.

4.1.1 Tweet Collection

Tweets are collected from well-known Arabic news accounts, which are: Al-Arabiya, Sabq, CNN Arabic, and BBC Arabic. These accounts belong to television channels and online newspapers, where they use Twitter to broadcast news related to real-world events. The first collection process tracks tweets from the news accounts for 20 days period, between November 2, 2021, and November 22, 2021 and we call this dataset AraEvent(November). We also pull test-specific data from different times to minimize the impact of bias due to the period in which we collected the data and we name it AraEvent(July). The retrieval process of the AraEvent(July) covers 6 days between July 6, 2022, and July 12, 2022, from the same news accounts and utilizing Twitter Streaming API [3]. As a first pre-processing step, each retweet by these accounts is filtered and excluded during the collection process. The AraEvent(November) and AraEvent(July) tweets datasets consist of around $12,095$ and 813 tweets, respectively. To ensure the quality of the datasets, text normalization is applied to convert tweets to a more standard form by eliminating noise from the data. The tweet normalization process removes the following: diacritics, punctuation marks, emoticons, URLs, user men-

[3] https://developer.twitter.com/en/docs/twitter-api

tions, emails. In addition, different surface forms of character Alif (آأإ) normalized to plain Alif (ا), and Taa Marbouta (ة) normalized to Haa (ه). Furthermore, we constrain the number of times the character is repeated to a maximum of two repetitions and replace successive spaces and newlines with one space separation. Non-Arabic characters are preserved as they might contain technical or scientific details related to the event. Hashtags are converted to plain text by separating the text into words as some hashtags may retain significant information about the event. However, Hashtags that contain the news account name in Arabic or English and also the Breaking hashtag both are removed because they are considered as redundant information, for example, #العربيه , #بي_بي_سي_ترندينغ, and #عاجل. Finally, tweets containing words less than 7 in total are filtered out along with duplicate tweets. Tables 5 and 6 in the Appendix show statistics for the final AraEvent(November) dataset and AraEvent(July) datasets, respectively, after the pre-processing steps conducted above.

4.1.2 Annotation

The annotation process is performed after the text normalization and filtering steps and is conducted manually by two of the authors and a volunteer, disagreement was resolved by discussion. To specify the annotation rules and conditions, we follow the "Automatic Content Extraction (ACE) Arabic annotation events guidelines" (ACE, 2005), ACE2005. Based on the definition from (ACE, 2005), an event is considered to be an action involving a connection between participants. Therefore, a special collection of events' types and subtypes are labeled and considered while annotating the events dataset. Accordingly, a set of unclassified news tweets written in Arabic are given, and after applying the annotation guidelines, the results are broken down into types and subtypes of events.

To start the annotation process, first, we consider and focus on eight event types mentioned in (ACE, 2005): "Life, Movement, Transaction, Business, Conflict, Contact, Personnel, and Justice events", and the corresponding sub-type for every main type. Tweet examples of types and sub-types of events following the (ACE, 2005), are presented in Table 10, Appendix A. Second, based on our data, we made some adjustments to the (ACE, 2005) guidelines to accommodate as many events as possible which are published on the Arabic news accounts. The modifications are either to expand the definition of a particular event type and make it include a larger segment of acts or to add a new subtype to the main event type. Furthermore, Table 11 in Appendix A summarizes the key modifications this study introduces, including defining the changes, why we perform them, and also present illustrative examples.

Additionally, the frequent occurrence of natural events, especially the natural disaster, in the dataset inspired us to propose a main event type, Nature, and a subtype of this event, namely Natural Disasters. Floods, earthquakes, volcanoes, pollution from volcanoes that cause a loss of life or property are labeled as natural disasters. The following tweet is an example of Nature type event with subtype as Natural-disaster:

- Arabic: "اطلق اعمده دخان وسحب رماد الى ارتفاع ٣٥٠٠ مترلحظه انفجار بركان جبل اسو في اليابان"
- Translation: "Smoke plumes and ash clouds were released to an altitude of 3,500 meters at the moment of the eruption of Mount Aso volcano in Japan"

In this work, we do not include fires as natural disasters due to the lack of information on the cause of the fire. There are also two types of events to consider: 'None' and 'Other'. The label 'None' is used if there is no identified event in the tweet. On the other hand, 'Other' type is used to label any event that is not from the pre-defined list of types that the system considers and if the constructions of the event are not clearly defined or ambiguous. Examples of tweets labeled in Twitter Event dataset with 'None' and 'Other' types respectively are:

- Arabic: "استشاري لهذا السبب العلاقه وثيقه بين سمنه الاطفال والميكروبات المعويه الضاره"
- Translation: "consultant, for this reason, the close relationship between childhood obesity and harmful intestinal microbes"
- Arabic: "الشريف القابضه تعلن شراكتها مع كيوبك ارت لانشاء محطات الشحن للسيارات الكهربائيه في المملكه"

Type1	Subtype1	Type2	Subtype2	Type3	Subtype3	Tweet
Life	Die	Life	Injure	Conflict	Attack	Arabic: الشرطه الاوغنديه مقتل ٣ اشخاص واصابه ٣٣ اخرين في هجومي كامبالا الانتحاريين Translation: "Uganda police 3 killed, 33 injured in Kampala suicide bombings"
Conflict	Attack	Life	Die	-	-	Arabic: في هجوم بالقوس والسهم مقتل عده اشخاص على يد مهاجم ب النرويج Translation: "In a bow and arrow attack, several people were killed by an attacker in Norway"

Table 2: Examples of more than on event type labels

- Translation:" Al-Sharif Holding announces its partnership with Cubic Art to establish charging stations for electric cars in the Kingdom"

A tweet can have one, two, or three main types associated with a subtype based on the occurrence of the event as the examples show in Table 2. As a consequence of the annotation approach described above, for the AraEvent(November) dataset, we end up with $2,146$ annotated events each with their corresponding type and subtype. In addition, 858 tweets contain 'Other' events, and a total of $8,069$ tweets are of the 'None' type. The AraEvent(July) dataset contains 110 annotated events, each with a type and subtype with 257 tweets of the 'Other' type and 446 of the 'None' type.

4.1.3 Annotation Results

In this section, we present the statistics of the types of events that exist on the AraEvent(November) and AraEvent(July) datasets at the level of one event or the set of events that took place simultaneously. The AraEvent(November) statistics are present in Appendix A Table 7 in terms of individual or paired events only. The individual event type with the highest frequency based on the data we have is the Justice event with 527 tweets. The Conflict type comes as the next highest event with 449 tweets. The Life type ranked third with 304 tweets. The least accounted event type in the data is the Transaction type with 41 tweets. Regarding the paired events, the two highest events that occur concurrently are Life and Conflict as they record 203 tweets. Second, Conflict and Justice events happen at once in 37 tweets. Nature and Life event types occurred 11 times as the third most overlapped event. On the other hand, Personnel and Business types overlap with one event type only as following: Personnel and Conflict, Justice and Business. In addition, a set of Life, Justice, and Conflict events occur at the same time in 9 tweets. Table 2 shows an example of Life and Conflict paired events. The following is an example of paired events between Nature and Life:

- Arabic: شعر به في السعوديه مصرع شخص واصابه ٣ اخرين في زلزال قوي جنوب ايران
- Translation: "It was felt in Saudi Arabia, one person was killed and 3 others injured in a strong earthquake in southern Iran."

Regarding the AraEvent(July) dataset, individual and paired event type statistics are present in Appendix A Table 8. In terms of the lowest individual event types recorded in the data are Transaction and Nature events with 2 tweets each. Moreover, no business events are accounted for in the data. The individual event type with the highest frequency, based on the data we have, is the Justice event with 24 tweets. With 23 tweets, the Life type is ranked second, and lastly, the Conflict type is ranked third with 20 tweets.

4.2 Evaluation Experimental Setup

The event detection problem is a *Multi-Label problem*. The same sentence can contain multiple events. We follow (Liu et al., 2019) approach, in which we convert the multi-label problem to *multiple binary classification* problems. As we have 9 event types, from Section 4.1, we fine-tuned *EBK-BERT* per event type. This fine-tuning is performed to the *RandMask* model, too. To evaluate the models, four experiments are conducted.

1. The first experiment aims to evaluate the models when applied to test data from the same duration. Train-test split is used with an 80:20 ratio of the AraEvent(November) dataset.

2. The second experiment aims to evaluate the models when applied to test data from the same duration, but with limited training samples. Training samples in this experiment were limited to 100 balanced samples, and testing varies between event types as it constituted the balanced remaining samples not consumed in training.

3. The third experiment aims to evaluate the models when applied to test data from a different

duration. The AraEvent(November) dataset is used for training, and AraEvent(July) dataset is used for testing. Business, Transaction, and Nature types were not considered in this experiment due to having less than 10 samples each.

4. The fourth experiment aims to evaluate the models when applied to test data from a different duration, and with limited training samples. This experiment is considered to be the strongest form of testing of the four setups. Training samples in this experiment are limited to 100 balanced samples from the AraEvent(November) dataset, and testing is done on the AraEvent(July) data.

In all the experiments, we balance the positive class of an event type with a mixture of the other 8 types that do not overlap with the positive class, in addition to sentences that do not contain any events. Therefore, the final dataset of an event type includes: 50% sentences from the positive class of the event, 25% sentences from the other event types, and 25% sentences that do not contain events, the total amount of records for each class is presented in Tables 3 and 4. To initiate the fine-tuning step, AutoModelForSequenceClassification class from the transformers library of Huggingface [4] is used. All models are fine-tuned on 3 epochs with a learning rate of $5e-5$, batch size of 8, and a maximum sequence length of 128. For evaluation, As the datasets are balanced, we only report the mean of the accuracy per event type with a confidence interval of 95%. The fine-tuning is repeated 10 times with random initial seeds.

5 Evaluation Results and Discussion

To evaluate the proposed approach, we compare between the classification results of the fine-tuning of both the baseline *RandMask* and our proposed approach *EBK-BERT*. Starting with the first and second experiment, as presented in Table 3, *EBK-BERT* performs better than *RandMask* in all types. The Business type had the most improvement with about 3.5% improvement in accuracy. Then comes Personnel, Movement, and Contact with $2-3\%$ improvement in accuracy. The remaining events show an improvement of less than $1.6-0\%$. When limiting the training data, the Business type still shows the highest improvement with 4.2%, The remaining types show an improvement of $0.4-3\%$ except for Nature, which is affected negatively by the EBK Masking. The average improvement is 2.13% and 1.4% for the two experiments respectively. We conclude from this that, for most of the types, EBK Masking did amplify the fine-tuning process to produce more accurate predictions for homogeneous datasets.

As for the third and fourth experiments, where the test set is from a different time period, the results are presented in Table 4. The average improvement of the third experiment is at 1.74% with 5 out of the 6 datasets scoring more than 1% improvement. The fourth experiment which limits the training size to 100 shows the promising results of EBK Masking when capturing the masks correctly. The average improvement from the EBK Masking is at 3.67%. However, this average comes from two opposite responses to EBK Masking when testing on non-homogeneous datasets. Conflict and Contact had an improvement of 0.6% and -0.9%, which is an indication of a bias in the selection of significant words which did not generalize well when tested in a different period with a different event. Emphasizing that the models perform much better when training on the entire training dataset. Whereas for the remaining four event types, more than $5-6.25\%$ improvement is archived by *EBK-BERT*. This indicates that *EBK-BERT* generalizes well for different time periods even with limited fine-tuning data. Still, it cannot be ascertained whether this is the reason for the varying performance between the types since there are a lot of variables that may play a role, such as the data size, the difficulty of the event, the bias in the most significant words, and the percentage of the presence of the most significant words in the pre-training text.

6 Limitations

AraEvent is drawn from a short period, introducing some bias towards events happening in that period such as راجحي and نيوكسل. Also, errors from the lemmatization tool propagate to the ED task, as shown in the Table 1 ترشح, رشح are both present in the most significant words. In addition, as the models were trained on MSA Arabic corpus, we cannot generalize the results to dialectal Arabic as it may impose its own challenges.

[4] https://huggingface.co

		80:20 training to test ratio			Training size set to 100 balanced samples		
Event type	**Training size**	**Testing size**	**Random**	**EBK-BERT**	**Testing size**	**Random**	**EBK-BERT**
Personnel	189	47	0.862±0.014	**0.891±0.013**	136	0.866±0.013	**0.862±0.013**
Transaction	71	17	0.733±0.039	**0.750±0.047**	-	-	-
Contact	347	86	0.934±0.007	**0.955±0.004**	333	0.903±0.006	0.913±0.008
Nature	123	30	0.917±0.010	0.923±0.012	53	**0.942±0.007**	0.930±0.010
Movement	148	37	0.858±0.009	**0.882±0.023**	85	0.800±0.023	**0.811±0.018**
Life	858	214	0.880±0.006	**0.917±0.003**	972	0.796±0.008	**0.825±0.009**
Justice	393	234	0.910±0.006	**0.927±0.003**	1073	0.798±0.012	**0.824±0.011**
Conflict	1134	283	0.897±0.002	0.904±0.005	1317	0.807±0.004	0.811±0.001
Business	103	25	0.869±0.023	**0.904±0.017**	28	0.911±0.016	**0.954±0.015**

Table 3: Event classification accuracy results for AraEvent(November) based on an average of 10 runs per event type and a confidence interval of 95%

	Testing Dataset	Full training set size			Training size set to 100 balanced samples	
Event type	**Testing size**	**Training size**	**Random**	**EBK-BERT**	**Random**	**EBK-BERT**
Personnel	**32**	236	0.913 ± 0.018	**0.93125 ± 0.020**	0.853± 0.024	**0.903 ± 0.006**
Contact	**21**	433	0.967 ± 0.014	0.962 ± 0.012	0.919 ± 0.020	0.910 ± 0.026
Movement	**24**	185	0.788 ± 0.0191	**0.804 ± 0.017**	0.746 ± 0.0284	**0.808 ± 0.030**
Life	**45**	1137	0.985 ± 0.006	**0.990 ± 0.003**	0.901 ± 0.0197	**0.952 ± 0.00762**
Justice	**23**	627	0.833 ± 0.016	**0.86 ± 0.0113**	0.769 ± 0.019	**0.829 ± 0.023**
Conflict	**39**	1417	0.827 ± 0.009	**0.869 ± 0.013**	0.770± 0.019	0.777 ± 0.017

Table 4: Event classification accuracy results of AraEvent(July) based on an average of 10 runs per event type and a confidence interval of 95%

7 Conclusion and Future Work

This work aims to propose using the Event-Based Knowledge *(EBK)* approach for selecting the Mask for the MLM training task in order to improve the model's performance for the event detection task. In this *(EBK)*, the most significant words are extracted from an AraEvent(November) using odds ratio. This dataset is pulled from news channels' Twitter accounts and then annotated manually to 9 event types, inspired by ACE2005 Event Extraction dataset, with some modifications. The event classification experiment results show improvement over random masking by $0.56 - 3.645\%$ across all event types when tested on a homogeneous dataset, an average of 3.67% when tested on a non-homogeneous dataset with limited fine-tuning data. This shows the effectiveness of the proposed masking technique for event detection. The classification results, although higher than random masking, raise several questions on the reasons for the varying performance across the types. Running the experiment with different data sizes may asssist to answer the question of whether the data size plays a role in narrowing the effect of the proposed masking, in other words: Is *(EBK)* or similar masking approaches more suitable to perform tasks with small annotated datasets? Another improvement to be made to the approach is constructing the event datasets gradually over an expanded period of time to mitigate the bias towards the data collection period. Also, following the log-odds-ratio with Dirichlet prior approach should help us mitigate the bias of the collection period and rare words, in general. We consider this study as preliminary work as a proof of concept that mask approaches catered to a certain downstream task are beneficial to the downstream task for language models built for the Arabic Language. This is illustrated in this study on a language model built on a considerably limited amount of data. It is interesting to see if this approach can be applied to pre-train large-scale Arabic Language models for different downstream tasks.

Acknowledgements

The authors would like to acknowledge the support of King Fahd University of Petroleum and Minerals to complete this work. We would like to thank also Reem Alyami and Ebtehal Alsulami for the valuable discussion and support.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California. Association for Computational Linguistics.

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training BERT on arabic tweets: Practical considerations. *CoRR*, abs/2102.10684.

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

ACE. 2005. (automatic content extraction) arabic annotation guidelines for events. *Linguistic Data Consortium*.

Alaa Alharbi and Mark Lee. 2021. Kawarith: an arabic twitter corpus for crisis events. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52.

Hind Almerekhi, Maram Hasanain, and Tamer Elsayed. 2016. Evetar: A new test collection for event detection in arabic tweets. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 689–692.

Nasser Alsaedi and Pete Burnap. 2015. Arabic event detection in social media. In *Computational Linguistics and Intelligent Text Processing*, pages 384–401, Cham. Springer International Publishing.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Tommaso Caselli, Osman Mutlu, Angelo Basile, and Ali Hürriyetoğlu. 2021. Protest-er: Retraining bert for protest event extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 12–19.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–419.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2018. A tf-idf and co-occurrence based approach for events extraction from arabic news corpus. In *International Conference on Applications of Natural Language to Information Systems*, pages 272–280. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *CoRR*, abs/1611.04033.

Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Mucahid Kutlu, and Hind Almerekhi. 2017. Evetar: Building a large-scale multi-task test collection over arabic tweets.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.

Shulin Liu, Yang Li, Feng Zhang, Tao Yang, and Xinpeng Zhou. 2019. Event detection without triggers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 735–744.

Shulin Liu, Kang Liu, Shizhu He, and Jun Zhao. 2016. A probabilistic soft logic based approach to exploiting latent and global information in event classification. In *AAAI*.

AS Mohammad and Omar Qawasmeh. 2016. Knowledge-based approach for event extraction from arabic tweets. *International Journal of Advanced Computer Science & Applications*, 1(7):483–490.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Mohammad Smadi and Omar Qawasmeh. 2018. A supervised machine learning approach for events extraction out of arabic tweets. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 114–119. IEEE.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

M. Szumilas. 2010. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*, 19(3):227–229.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019. Hmeae: Hierarchical modular event argument extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5777–5783.

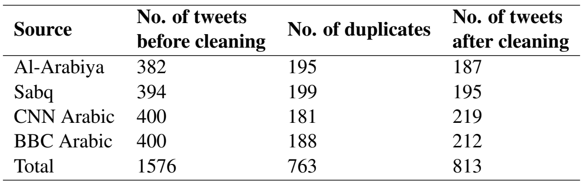
Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294.

Tianyi Zhang and Tatsunori Hashimoto. 2021. On the inductive bias of masked language modeling: From statistical to syntactic dependencies. *arXiv preprint arXiv:2104.05694*.

A Appendix

Source	No. of tweets before cleaning	No. of duplicates	No. of tweets after cleaning
Al-Arabiya	2945	170	2775
Sabq	2837	143	2694
CNN Arabic	3192	574	2618
BBC Arabic	3121	135	2986
Total	12095	1022	11073

Table 5: AraEvent(November) data statistics



Source	No. of tweets before cleaning	No. of duplicates	No. of tweets after cleaning
Al-Arabiya	382	195	187
Sabq	394	199	195
CNN Arabic	400	181	219
BBC Arabic	400	188	212
Total	1576	763	813

Table 6: AraEvent(July) data statistics

Type	Personnel	Transaction	Contact	Nature	Movement	Life	Justice	Conflict	Business
Personnel	116	-	-	-	-	-	-	1	-
Transaction	-	41	-	-	1	1	1	-	-
Contact	-	-	216	-	1	-	-	-	-
Nature	-	-	-	60	1	11	1	4	-
Movement	-	1	1	1	80	1	3	6	-
Life	-	1	-	11	1	304	7	203	-
Justice	-	1	-	1	3	7	527	37	2
Conflict	2	-	-	4	6	203	37	449	-
Business	-	-	-	-	-	-	2	-	62
Total	118	44	217	77	93	527	578	700	64

Table 7: AraEvent(November) event types statistics

Type	Personnel	Transaction	Contact	Nature	Movement	Life	Justice	Conflict	Business
Personnel	16	-	-	-	-	-	-	-	-
Transaction	-	2	-	-	-	-	-	-	-
Contact	-	-	11	-	-	-	-	-	-
Nature	-	-	-	2	-	-	-	-	-
Movement	-	-	-	-	12	-	-	1	-
Life	-	-	-	-	-	23	-	7	-
Justice	-	-	-	-	-	-	24	-	-
Conflict	-	-	-	-	1	7	-	20	-
Business	-	-	-	-	-	-		-	-
Total	16	2	11	2	13	30	24	28	-

Table 8: AraEvent(July) event types statistics

	$\|\|words\|\|$ before lemmatization	Odds ratio >2 (Before lemmatization)	$\|\|words\|\|$ after lemmatization	Odds ratio >2 (After lemmatization)	Presence of the words in the pre-training corpus
Personnel	1017	445	798	272	11.40%
Transaction	470	283	404	375	8.31%
Contact	1328	447	975	447	14.69%
Nature	519	260	415	237	11.57%
Movement	766	413	637	363	9.72%
Life	2890	347	1922	305	6.08%
Justice	3619	371	2316	353	7.94%
Conflict	3700	455	2232	393	7.68%
business	648	303	532	289	10.48%
Total	14957	3324	10231	3034	-

Table 9: Statistics related to the significant words calculated by the odds-ratio

Type	Sub$_{type}$	Tweet
Life	Be-Born	Arabic: في مثل هذا اليوم ولد احد اشهر ادباء روسيا والتاريخ هل قرات سيرته او بعضا من اعماله Translation:" On this day, one of the most famous writers of Russia and history was born. Have you read his biography or some of his works"
	Marry	Arabic: اميره يابانيه تتزوج من صديقها وتتنازل عن صفتها الملكيه Translation:"Japanese princess marries her boyfriend and relinquishes her royal status"
	Divorce	Arabic: علمت من مواقع التواصل فاتن موسى تكشف تفاصيل صادمه عن طلاقها من الفنان مصطفى فهمي Translation:"She knew from social media Faten Moussa reveals shocking details about her divorce from the artist Mustafa Fahmy"
	Injure	Arabic: المنطقه الخضراء في بغداد اصابه ١٢٥ شخصا من القوات الامنيه والمتظاهرين المحتجين على نتائج الانتخابات Translation:"Baghdad'S Green Zone injured 125 security forces and protesters protesting the election results"
	Die	Arabic: قتل مالكولم اكس بالرصاص في قاعه رقص في مدينه نيويورك امام عائلته قبل ٥٦ عاما وكان يبلغ وقتها من العمر ٣٩ عاما Translation:"Malcolm X has shot dead in a New York City dance floor in front of his family 56 years ago at the time at the age of 39"
Movement	Transport	Arabic: الامير عبدالعزيز بن سعود يصل مملكه البحرين Translation:"Prince Abdulaziz bin Saud arrives in Bahrain"
Transaction	Transfer-Ownership	Arabic: السعوديه صندوق الاستثمارات السيادي يعلن استحواذه على ١٠٠ بالمائه من نادي نيوكاسل يونايتد Translation:"Saudi Sovereign Investment Fund announces acquisition of 100 percent of Newcastle United club"
	Transfer-Money	Arabic: الربيعه اكثر من مليار و٩٥٨ مليون دولار حجم المساعدات المقدمه من المملكه لبرنامج الاغذيه العالمي Translation:"Al- Rabiah more than one billion and 958 million dollars in aid from the Kingdom to the world food program"
Business	Start-Org	Arabic: اسواق عبدالله العثيم تفتتح احدث فروعها بحي النسيم ب سيهات Translation:"Abdullah Al-Othaim Markets opens its newest branches in Al-Naseem neighborhood in Sihat"
	End-Org	Arabic: امانه العاصمه المقدسه تغلق ٦ منشات تجاريه مخالفه للانظمه البلديه Translation:"Holy capital municipality closes 6 commercial establishments contrary to municipal regulations"
Conflict	Demonstrate	Arabic: احتجاجات في اصفهان على جفاف نهر زاينده بعد تحويل مجراه والسلطات الايرانيه تقطع الانترنت عن المنطقه Translation:"Protests in Isfahan over the drought of the Ziande River after the diversion of its course and the Iranian authorities cut off the Internet from the region."
	Attack	Arabic: ارمينيا و اذربيجان تتبادلان اطلاق النار على الحدود قرب اقليم ناغورنو كاراباخالذي شهد العام الماضي حربا بين هذين البلدين وكل طرف يتهم الثاني بالتسبب بالواقعه Translation:"Armenia and Azerbaijan exchanged fire on the border near Nagorno-Karabakh, which last year witnessed a war between these two countries and each side accuses the second of causing the incident."
Contact	Meet	Arabic: الاسد التقى بوزير الخارجيه الاماراتي في دمشق Translation:"Assad met with UAE Foreign Minister in Damascus"
	Phone-Write	Arabic: الامير محمد بن سلمان يجري اتصالا هاتفيا للاطمئنان على صحه رئيس وزراء العراق مصطفى الكاظمي Translation:"Prince Mohammed bin Salman makes a phone call to check on the health of Iraqi Prime Minister Mustafa Al-Kadhimi"
Personnel	Start-Position	Arabic: اللبنانيه ساره منقاره تنضم لاداره بايدن كمستشاره خاصه لشؤون حقوق ذوي الاحتياجات الخاصه وهذا ما قالته ل حول ذلك Translation:"Lebanese Sarah Manqara joins Biden'S administration as a special adviser on special needs rights and that's what she said about that"
	End-Position	Arabic: المجلس الرئاسي الليبي يوقف وزيره الخارجيه نجلاء المنقوش عن العمل ويمنعها من السفر Translation:"Libyan Presidential Council suspends foreign minister Najla alManoush from work and prevents her from traveling"
	Nominate	Arabic: مرشح الجزائر محمد هامل امينا عاما جديدا لمنتدى الدول المصدره للغاز Translation:"Algeria's Candidate Mohamed Hamel as New Secretary-General of the Forum of Gas Exporters"
Justice	Arrest-Jail	Arabic: السلطات التركيه تعتقل العشرات بعد احتجاجات عنيفه ضد اللاجئين السوريين في انقره Translation:"Turkish authorities arrest dozens after violent protests against Syrian refugees in Ankara"
	Release-Parole	Arabic: مصر الافراج عن الناشطه اسراء عبد الفتاح Translation:"Egypt releases activist Israa Abdel Fattah"
	Trial-Hearing	Arabic: بدء محاكمه المغني الامريكي كيلي الحائز على جائزه غرامي بتهم ممارسه الابتزاز وتضليل العداله والاعتداء الجنسي على قاصرات Translation:"American Grammy award-winning singer Kelly begins trial on charges of abusing, misleading justice, and sexually assaulting minors"
	Charge-Indict	Arabic: على خلفيه مخطط لاغتيال مواطنين اسرائيليين في الجزيره السلطات القبرصيه توجه اتهامات جنائيه لسته اشخاص Translation:"Regarding assassinate Israeli citizens on the island, the Cypriot authorities are pressing criminal charges against six people"
	Sue	Arabic: فتاه ترفع دعوى في الولايات المتحده على الامير اندرو بتهمه الاعتداء الجنسي Translation:"Girl sues Prince Andrew in U.S. for sexual assault"
	Convict	Arabic: النيابه العامه صدور حكم ابتدائي بادانه احد المتهمين ظهر في مقطع فيديو يتضمن قيامه بالتحرش بامراه Translation:"The public prosecution issued a preliminary verdict in the conviction of one of the accused appeared in a video clip that includes him harassing a woman"
	Sentence	Arabic: محكمه تركيه تقضي بسجن زوجه رئيس حزب معارض ل اردوغان عامين ونصف بسبب تقرير طبي Translation:" Turkish court sentences wife of the president of the opposition party to Erdogan with two and a half years in prison for medical report"
	Fine	Arabic: الف دولار قيمه الغرامات بحق نائبه رفضت ارتداء الكمامه بمبنى الكونغرس ٤٨ Translation:"$48, 000 worth of fines for a deputy who refused to wear the mask in the Capitol"
	Execute	Arabic: الاعدام لقاتل مدير بلديه كربلاء واستمرار الجدل على ادانه عناصر حمايته Translation:" Execution of the killer of the mayor of Karbala and the continuing controversy over the condemnation of his protection elements"
	Extradite	Arabic: اليكس صعب تسليم المساعد البارز لرئيس فنزويلا نيكولاس مادورو الى الولايات المتحده Translation:"Alex Saab delivers Venezuela's top aide Nicolas Maduro to the United States"
	Acquit	Arabic: انهار امام المحكمه واجهش في البكاء القضاء الاميركي يبرئ الاميركي كايل ريتنهاوس من جريمه قتل Translation:"He collapsed in front of the court and faced in tears the American judiciary cleared American Kyle Rittenhouse of murder"
	Appeal	موفد الى انتويرب نور الدين الفريضي محكمه بلجيكيه تنظر الاستئناف المقدم من خليه اسد الله اسديواحد المتهمين يتعاون مع المحققين Translation:"A Belgian court is hearing the appeal filed by Assadullah Asadi'S cell, one of the accused is cooperating with the investigators"
	Pardon	Arabic: ملك الاردن يعفو عن ١٥٥ محكوما باطاله اللسان عليه Translation:"King of Jordan pardons 155 convicted of Offensive Speech"

Table 10: Examples of Labeling Results Following Guidelines from (ACE, 2005).

Type		Subtype			
Name	**Changes from ACE2005 Description if any**	**Name**	**Exist in ACE2005**	**Reason for Adding/changing**	**Example**
Business	No changes	Buy	No	Transfer-Ownership sub-type restricted the buying and selling events for artifacts such as vehicles or weapons and organizations or facilities. Thus, we label the event as Buy or Sell when any physical item is purchased or sold respectively. (excluding items from Transfer-Ownership).	Arabic: تهافت الدول الغنية على شراء سلاح كورونا الجديد هل يصل إلى الفقيرة منها Translation:"Rich countries are eager to buy a new corona weapon will it reach the poor ones?
		Sell	No		Arabic: لوحه بانكسي المزقه تباع مجددا برقم قياسي قدره ٢٥٤ مليون دولار Translation:"Banksy torn painting is selling again for a record $254 million"
Justice	In contrast with ACE2005, we label the JUSTICE event subtypes based on actions by any entity that holds authority or dominance and not only the government, because the governing authority in a state that is experiencing aggression and wars might not have the authority over the conflict zones where the event occurs or it might be an occupation authority.	Accusation	No	The Accusation event occurs when a person, country, entity, president, or government speaker claims and accuses another state, person, or entity without evidence or a trial. In contrast with the Charge-Indict sub-type, which happen when there's a person or organization accused of a crime by a government actor.	Arabic: اثينا تهم انقره بتوجيه زورق مهاجرين الى المياه اليونانيه Translation:"Athens accuses Ankara of directing migrant boats into Greek waters"
		Condemnation	No	The Condemnation event takes place when a state, official, organization, or persons condemn an act or attack by another country or organization , or people without any trial. In contrast with the Convict sub-type which occurs when the court conviction the accused.	Arabic: الخارجيه الالمانيه تدين استيلاء الحوثي على السفاره الاميركيه في صنعاء Translation:"German Foreign Affairs condemns the Houthi takeover of the US embassy in Sana'a.
Conflict	No changes	Attack	Yes	The attack sub-type event will be expanded from its original definition which includes "Conflict, clashes, fighting, and shooting" to cover other important attacks such as sexual harassment, rape, burglary, Cyberattack, and takeover a facility.	Arabic: عصابه برازيليه تسطو على بنك وتهرب بالرهائن مقيدين فوق اسطح السيارات Translation:"Brazilian gang robs a bank and escapes hostages tied up on car roofs"
Movement	No changes	Transport	Yes	The sub-type event Transport will expand to cover the transportation of physical items and not the items will not be limited to a weapon, vehicle, or even a person as defined in ACE2005.	Arabic: مركب الشمس نقل مركب الفرعون خوفو الى المتحف المصري الكبير Translation: "The sun boat transported Pharaoh Khufu's boat to the Great Egyptian Museum"
Contact	No changes	Meet	Yes	We modified the definition of the sub-type event Meet to happen when two or more entities meet and interact with each other regardless of whether they're in the same location or not which is the constraint that was identified by ACE2005.	Arabic: لبحث سبل اداره التنافس بين البلدين بشكل مسؤولالبيت الابيض بايدن ونظيره الصيني يعقدان اجتماعا افتراضيا الاثنين المقبل Translation: "White House Biden and His Chinese Counterpart Hold Virtual Meeting next Monday to discuss ways to manage competition between the two countries responsibly"
Personnel	No changes	Nominate	Yes	The event will occur when a person has run or become a candidate in a race for either a party or presidential nomination and is not limited to a proposed person for a specific position by organizations.	Arabic: الدبيبه يرشح نفسه رسميا للانتخابات الرئاسيه الليبيه Translation:"Aldebiba officially nominates himself for Libyan Presidential elections "

Table 11: Examples of Labeling Results After modifying the Guidelines