Detecting and Geocoding Battle Events from Social Media Messages on the Russo-Ukrainian War: Shared Task 2, CASE 2023

Hristo Tanev
Joint Research Centre
European Commission
Ispra, Italy
hristo.tanev
@ec.europa.eu

Nicolas Stefanovitch
Joint Research Centre
European Commission
Ispra, Italy
nicolas.stefanovitch
@ec.europa.eu

Andrew Halterman
Michigan State University
Department of Political Science
Michigan, USA
ahalterman0@gmail.com

Onur Uca
Sociology Department
Mersin University
Mersin, Turkey
onuruca@mersin.edu.tr

Vanni Zavarella
University of Cagliari
Cagliari, Italy
v.zavarella@unica.it

Ali Hüriyetoğlu
KNAW Humanities
Cluster DHLab
Netherlands
ali.hurriyetoglu
@dh.huc.knaw.nl

Bertrand De Longueville
Joint Research Centre
European Commission
Ispra, Italy
bertrand.de-longueville
@ec.europa.eu

Leonida Della Rocca
Engineering S.p.A.
Rome, Italy
leonida.della-rocca
@ext.jrc.ec.europa.eu

Abstract

The purpose of the shared task 2 at the Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) 2023 workshop was to test the abilities of the participating models and systems to detect and geocode armed conflicts events in social media messages from Telegram channels reporting on the Russo Ukrainian war. The evaluation followed an approach which was introduced in CASE 2021 (Giorgi et al., 2021): For each system we consider the correlation of the spatio-temporal distribution of its detected events and the events identified for the same period in the ACLED (Armed Conflict Location and Event Data Project) database (Raleigh et al., 2010). We use ACLED for the ground truth, since it is a well established standard in the field of event extraction and political trend analysis, which relies on human annotators for the encoding of security events using a fine grained taxonomy. Two systems participated in this shared task, we report in this paper on both the shared task and the participating systems.

1 Introduction

Automatic discovery of an event's location is an important sub-task of event extraction: most events occur at a defined location, reported in text. Usually the event time can be guessed by the time of the publication of the news article or the social media post and the presence of temporal adverbs. However, it is far more difficult to detect the location: multiple events can be reported in the same story, each with potentially no, one, or multiple locations mentioned in the text (Halterman, 2019; Radford, 2021; Akdemir et al., 2018).

Event geoparsing, as distinguished from simple geoparsing, is an important part of the event extraction process (Halterman, 2019; Dewandaru et al., 2020; Halterman, 2023). The purpose of this shared task was to provide a real-world evaluation of event geoparsing and challenge the researchers, working on event detection, to propose solutions for event geocoding. Another critical aspect of this evaluation is the comparison between automated and manually curated datasets in line with Giorgi

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 160–166, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_021

et al. (2021) and Zavarella et al. (2022).

Our evaluation methodology is based on spatio-temporal correlation, using the PRIO GRID geographical cells (Tollefsen et al., 2012): We measured the correlation between the geographical cells in which armed clashes were detected by the participating systems and the cells containing events from the gold standard data. Details about the evaluation methodology are given in the section *Data set and evaluation methodology*.

In the previous two years the shared task has featured protest events with complex geographical patterns. This year data, referring to Russo Ukrainian conflict, features battles situated along the Russian Ukrainian border.

Conflict has a different structure than protest. Protests are followed instantly by journalists, there is a civilian population, you can get information about the same protest from different news sources. In a military conflict it is difficult to access information as there is much less reporting from open source. And the information is often unreliable and imprecise. Conflict or their shape and size can be hidden or difficult to assess. All these are the main reasons why this work is both valuable and difficult.

This year we had two submissions, which used two different paradigms to event detection, exhibiting different behaviour: The TMA system, a combination of transformer-based classification model and a geoparser, which achieved better correlation and NEXUS, a rule based system also combined with a geoparser.

2 Related work

Socio-political event extraction (SPE) has long been a challenge for the natural language processing (NLP) community, as reflected in previous editions of the Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) workshops (Hürriyetoğlu et al., 2022). Specifically, event extraction in the security domain has been identified as an important application area in the automatic information retrieval domain (Best et al., 2008). Similarly, deriving geolocated information from social networks has been seen early identified as an application-rich discipline (Intagorn et al., 2010; De Longueville et al., 2010). Despite the fact that detection and geocoding of events from social media sources have been studied for more than a decade, the field is still vibrant and innovative as advances in Artificial Intelligence make new approaches possible, and as the evolution of the Web and its social media services constitute a "moving target" for automatic information extraction efforts.

3 Data

The goal of this task is to evaluate the performance of automatic discovery of event locations systems on modeling the spatial and temporal patterns of violence in the Russo-Ukrainian War. The data consists of Telegram messages from channels reporting about developments in the Russo-Ukrainian war. We evaluate the capability of participant systems to reproduce the manually curated Russo-Ukrainian War-related dataset.

3.1 Input Data

We provided one collection of English-language messages from Telegram channels with a large number of followers and constant broadcasts about the Russo-Ukraine war. The data was scraped using the official API from Telegram.

Telegram Telegram is the most important social media of data for this topic as it is very popular in the belligerant countries: Russia ranks second in the world in terms of Telegram users (24.15 million) and Ukraine ranks eighth (7.02 million). Data was scraped from Russian and Ukrainian Telegram accounts with a large number of followers who posted messages in English using the official Telegram API. We gathered nearly 326K original English Telegram Massages from Telegram Channels. Table 1 shows the Telegram Channels used and the number of followers.

Intel Slava Z	418 374 subscribers
MoD Russia	95 788 subscribers
Ukraine NOW	157 719 subscribers
Pravda_Gerashchenko_en	24 003 subscribers
UKR LEAKS_eng	52 226 subscribers
Ukraine Today	21 545 subscribers

Figure 1: Telegram Channels (verified channels) - (English Language)

The date ranges of the Telegram data and the date ranges of the gold standard are the same. The date range of Telegram data is February 24, 2022 / August 24, 2022

3.2 Gold Standard Data

The Armed Conflict Location and Event Data Project (ACLED) collects real-time data on the lo-

cations, dates, actors, fatalities, and types of all reported political violence and protest events around the world. The data ACLED collects is detailed and manually curated. For this study, we have used ACLED data from the date range: February 24, 2022 / August 24, 2022, and considered as events only the events located in Ukraine with the `Battle` event type. After the specified edits, we have an ACLED data set of 18K rows. This dataset was used as the gold standard data for the study.

We challenged the participant systems to reproduce the Gold Standard data set from ACLED's Curated Data comprising curated disorder events directly related to the Russo-Ukrainian War.

4 Evaluation

The performance of event geolocation is evaluated by computing correlation coefficients on event counts aggregated on cell-days, using uniform grid cells of approximately 55 kilometers sides from the PRIO-GRID data set (Tollefsen et al., 2012). We use these analytical measures as a proxy to the spatio-temporal pattern of violence in the Russo-Ukrainian War.

4.1 Metrics

We use the cell-days counts for two different analysis: the correlation with the total daily "Battle cell" counts (i.e., time trends alone) and the event counts for each cell-day (i.e., spatial and temporal trends together).

Temporal Trends The first analysis only considers the total number of "activated" cells (i.e., for which at least one Battle typed event was recorded), in the system output and Gold Standard data set. This time series analysis is sufficient to estimate how well the automatic systems capture the time trends of the conflict. However, it does not compute accuracy of system data in estimating the spatial variation of the target process.

Spatial and Temporal Trends We also measure the correlation coefficients on the absolute event counts with respect to Gold Standard, over each single cell-day.

For both analyses, we use two types of correlation coefficients to assess variable's relationship: Pearson coefficient r and Spearman's rank correlation coefficient ρ. Moreover, we used Root Mean Squared Error (RMSE) to measure the absolute value of the error on estimating cell/event counts from the Gold Standard.

5 Participating systems

5.1 XLM-RoBERTa and NEROne

The **TMA** system was composed of two modules: event classification and geolocation. The classifier was a `xml-roberta-small` (Liu et al., 2019) transformer model fine tuned using data from the ACLED dataset on all the 26 fine-grained classes using a batch size of 32 and 3 epochs. The training data was sampled over several years over 800k availlable data point in such a way to avoid highly skewed distribution: a maximum of 1k data points for each category, which resulted in a relatively small dataset of 23.6k datapoints and also lead to using the small version of the model instead of the large one.

The geolocation was performed using the NEROne system (Steinberger and Pouliquen, 2007) which is mulitlingual system based on the geonames dataset[1] with flexible matching and linking capacities, and which is able to provide the 3-levels of geographical information as expected by the scorer. Moreover, NEROne is able to guess the most likely place name among all the different geographical entities mentioned in a text.

An event was reported for a given text only if the ACLED type matched any label under `Battle` event type, and if a most likely place name was identified and it was located in Ukraine, moreover only entities for which the 3 levels of geolocation were predicted were considered. NEROne has the possibility to detect time expressions in a text, whenever that was the case, the date reported by NEROne was used, otherwise the publication date was considered.

5.2 NEXUS and Mordecai3

NEXUS is a multilingual event extraction system (Tanev et al., 2008) in the domain of conflict and disasters. It exploits language resources which are learned semi-automatically (Tanev et al., 2009). NEXUS is running as a module inside the Europe Media Monitor (EMM) (Best et al., 2005). In this shared task, however, we have run NEXUS as a standalone system, in order to discovers armed conflicts, reported in these posts. Regarding the spatio-temporal components of the detected events, NEXUS uses as event time, the time when the post

[1] http://www.geonames.org/

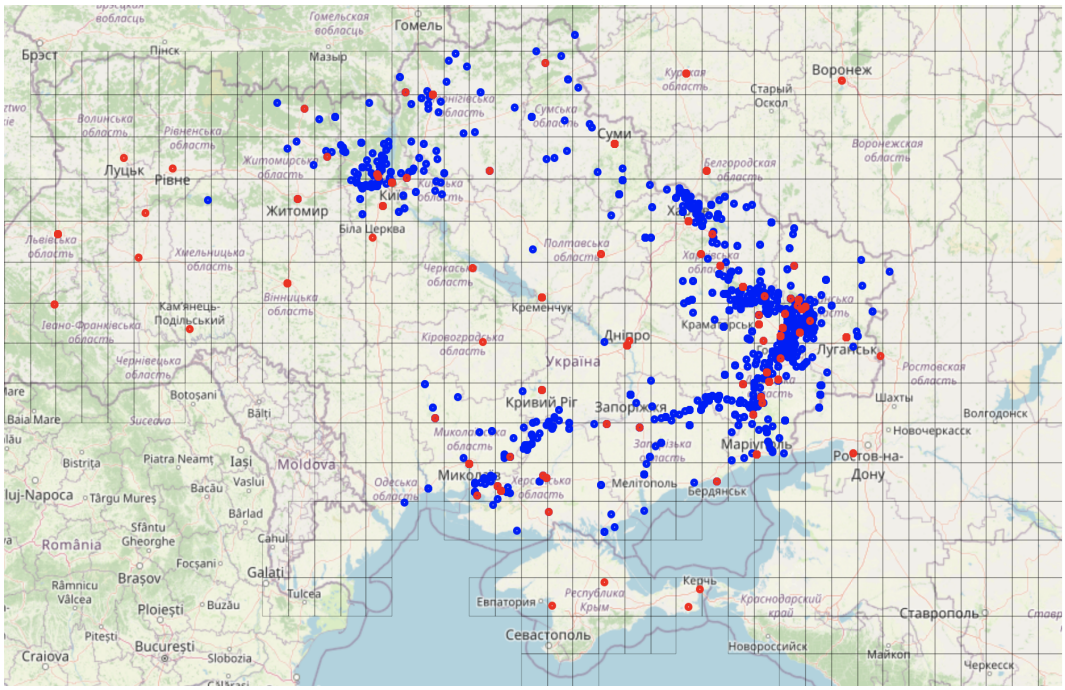

Figure 2: The geo-referenced Ukraine-Russo conflict records from Gold Standard (small blue dots) overlaid with the PRIO-GRID cells over the Ukraine.The red dots represent events recognized by the XLM-RoBERTa classification model and NERone system from Telegram.

was published, while the location is detected with the Mordecai3 geoparser (Halterman, 2023).

NEXUS classifies news articles and social media posts into a taxonomy of security related events, disasters, and humanitarian crises. Among the security related event classes, the system is capable of detecting military events, such as *battles*, *air attacks* and *shelling*, *criminal events*, such as robbery, kidnapping, murder, rape, assault, cyberattacks, as well as legal events such as trial and arrest.

Apart from the event type, location and time, NEXUS also detects other event metadata, such as conflict and crime perpetrators, dead and injured victims, kidnapped people, arrested, and displaced during war and disaster. Figure 3 shows an overview of the NEXUS event template.

Event classification is performed through AND/OR combinations of keywords, learned through weakly supervised multilingual terminology learning (Tanev, 2022). For the English language NEXUS uses a statistical SVM classifier, whose output is combined with the keyword classification, using empirically derived heuristics.

For our shared task run we filtered only the news which contain events of type *Armed conflict*, which is the NEXUS equivalent of the ACLED *Battle*.

Mordecai3 (Halterman, 2023) is an event geoparser that employs a two-step process for identi-

$$\begin{bmatrix} Date \\ Location \\ EventType \\ Dead \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Injured \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Kidnapped \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Displaced \begin{bmatrix} Number \\ Description \end{bmatrix} \\ Arrested [Description] \\ Perpetrators [Description] \\ Weapons \end{bmatrix}$$

Figure 3: Event template generated by the NEXUS event extraction system

fying an event's locations and resolving them to their geographic coordinates. First, it identifies all place names in the input text using named entity recognition and attempts to resolve each to their entry in the Geonames gazetteer (Wick and Boutreux, 2011). As features, it uses string and vector similarity between the extracted placenames and candidate geolocations from the Geonames gazetteer, along with contextual information from the other placenames present in the text. It uses these features in a neural networked trained on several thousand labeled events to select the best entry from Geonames. To conduct the second step of linking events and

	r	ρ	RMSE
NexMor3	0.127	0.155	98.70
TMA	0.338	0.295	73.40

Table 1: Correlation coefficients and error rates for daily Battle cell counts: r represents Pearson correlation coefficient, ρ is Spearman's rank correlation coefficient, and RMSE is the Root Mean Squared Error computed on day-cell units.

locations, it uses a fine-tuned question answering model (Halterman et al., 2023) that asks variations of "Where did [event] take place?" and identifies the location names that overlap with the answer span. Mordecai3 can identify multiple locations for a single event if they are present.

Only Telegram documents with ArmedConflict events (the NEXUS' counterpart of ACLED's Battle) identified with NEXUS were processed with Mordecai3.

6 Results

Table 1 shows the Pearson *r*, Spearman correlation coefficient ρ and Root Mean Squared Error (RMSE) of the total daily "Battle cell" counts of the two participant systems with respect to the Gold Standard, over the 6 months target time range. Here, the correlations are between the total number of cells per day where the system found an event vs. the number of cells where an event happened according to the Gold Standard (i.e., temporal patterns and not spatial patterns). These correlation measures are tolerant to errors in geocoding (as long as the events are located in Ukraine) and estimate the capability of the systems to detect from the source texts the evolution over time of the military clash events, independent of their location. We see that TMA system largely outperforms the Nexus-Mordecai3 system (*NexMor3* in the table) in both Pearson r and Spearman ρ coefficients.

Table 2 reports Pearson *r*, Spearman correlation coefficient ρ, and Root Mean Squared Error (RMSE) over cell-day event counts of the two participant systems with respect to Gold Standard, for the 6 months time range. Here the variables range over the whole set of PRIO-GRID cells included in the Ukraine territory and, thus, shows the correlation of event numbers across geo-cells, thus evaluating the systems' geolocation capabilities. The correlation scores for this metrics are in the lower to insignificant range as well for both systems, with a noticeable prevalence of TMA over Nexus-Mordecai3.

	r	ρ	RMSE
NexMor3	0.083	0.088	0.002
TMA	0.180	0.196	0.002

Table 2: Correlation coefficients and error rates for *cell-day* event counts of the Baseline and participant systems with respect to Gold Standard.

In Figure 4 and 5 we plot the time series of total daily Battle cells for the Gold Standard and TMA and Nexus-Mordecai3 systems, respectively. Only the TMA system seems to slightly capture the variation in the temporal pattern (i.e., an initial large number of Battle events which gradually declines, with recurrent escalations), but both system systems detect only a fraction of the events: While the average number of event per day is ca 10, the average number of event detected by the TMA system is around 2.5.

A more lenient representation of the agreement with Gold Standard is shown in Table 3. Here we report the confusion matrix between grid cells that Gold Standard and system runs code as experiencing at least a Battle event. It can be observed that only few of the cells classified as Battle by Gold Standard are detected by the automatic systems, which on the other hand incorrectly classified as Battle several additional cells.

6.1 Discusion

The correlations with the Gold standard obtained by both systems in this year shared tasks were much lower than the performance of the systems in the 2021 issue of same task, when data from the Black Lives Matter protests (Giorgi et al., 2021) were used as a Gold standard. Moreover, the Nexus system was also used in this 2021 shared task issue, achieving six times higher temporal correlation with the Gold standard than on the data from Russo Ukrainian conflict. This clearly shows that detecting and geolocating battles from the Russo Ukrainian war was far more challenging than replicating the data from Black Lives Matter protests. Table 3 shows that both systems have very low recall 2% and 9.3% and overall poor performances. There are several potential reasons for could lead to these results outside the intrinsic performance of each system: a) it could be that the data sample from Telegram channel did not contain the actual information allowing to recover the information present in the ACLED dataset; b) it could be that the data is unverified or biased, as such the systems are penalized even if the correctly detect the event

		Gold Standard true	Gold Standard false	Precision	Recall	F1
TMA	true	157	220	0.416	0.093	0.152
	false	1530	2435255			
NexMor3	true	39	75	0.34	0.02	0.04
	false	1648	2435400			

Table 3: Confusion matrix of grid cells experiencing at least one Battle event (true) versus inactive cells (false), for the Gold Standard and the participant systems.

and the location contained in a message. Properly assessing these will require further research.

The TMA system performs better at event classification, this could be due to the fact that it is a state of the art transformer-based model, but also the fact that it was trained on ACLED data, therefore having trained to detect the very types in the ground trurth could also play a role. It is not possible to assess properly which geoparser was the most efficient as the correlation as reported location depend on detected events.

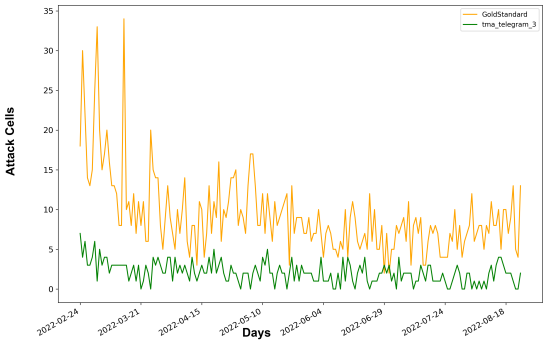

Figure 4: Time series of total daily Battle cells from the Gold Standard (in yellow) against TMA XLM RoBERTa/NERone runs on Telegram input data (in green).

7 Conclusions

The purpose of the database replication shared task is to provide a flexible benchmark for evaluation and comparison between event geocoding systems without annotated corpus of events and locations.

This year we tested the capabilities of the event detection systems to detect and geolocate battles event type in the Russo-Ukrainian war from Telegram messages in English, comparing the extracted events against a subset of the ACLED database, dedicated to the war in Ukraine. Two systems participated this year: Each system was an aggregation of two subsystems - event detection and classification and a geoparser, based on different paradigms.

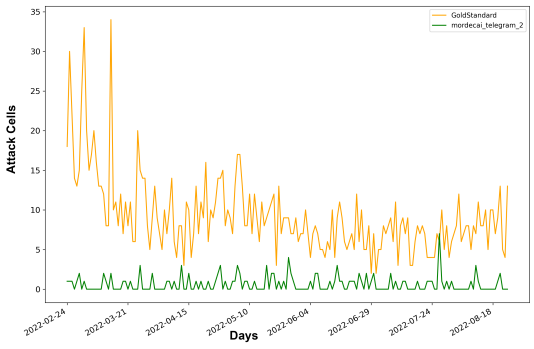

Figure 5: Time series of total daily Battle cells from the Gold Standard (in yellow) against NEXUS-Mordecai3 system runs on Telegram input data (in green).

The first system was a combination of Nexus and the Mordecai3 geoparser and the second consisted of event classifier based on XLM-RoBERTa combined with NERone geoparser. XLM RoBERTa and NERone obtained much better correlation in both evaluation scenarios: temporal and spaciotemporal.

A conclusion from this year shared task is that tracking armed conflicts is a challenging task, due to the incompleteness of the information: biased because of political consideration or unavailable because of security reasons, and in most case difficult to verify. Nevertheless, one of the participating systems achieved a medium level of correlation, which is a satisfactory result, given the difficulty of this year task.

References

Arda Akdemir, Ali Hürriyetoğlu, Erdem Yörük, Burak Gürel, Çağri Yoltar, and Deniz Yüret. 2018. Towards generalizable place name recognition systems: Analysis and enhancement of ner systems on english news from india. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, GIR'18, New York, NY, USA. Association for Computing Machinery.

Clive Best, Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, and Hristo Tanev. 2008. Automating event extraction for the security domain. In *Intelligence and Security Informatics*.

Clive Best, Erik van der Goot, Ken Blackler, Teófilo Garcia, and David Horby. 2005. Europe media monitor. *Technical Report EUR221 73 EN, European Commission*.

Bertrand De Longueville, Gianluca Luraschi, Paul Smits, Stephen Peedell, and Tom De Groeve. 2010. Citizens as sensors for natural hazards: A vgi integration workflow. *Geomatica*, 64(1):41–59.

Agung Dewandaru, Dwi Hendratmo Widyantoro, and Saiful Akbar. 2020. Event geoparser with pseudo-location entity identification and numerical argument extraction implementation and evaluation in indonesian news domain. *ISPRS International Journal of Geo-Information*, 9(12):712.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, Martin Andrews, Tiancheng Hu, Niklas Stoehr, Francesco Ignazio Re, Daniel Vegh, Dennis Atzenhofer, Brenda Curtis, and Ali Hürriyetoğlu. 2021. Discovering black lives matter events in the United States: Shared task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227, Online. Association for Computational Linguistics.

Andrew Halterman. 2019. Geolocating political events in text. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 29–39.

Andrew Halterman. 2023. Mordecai3: A neural geoparser and event geolocator. *working paper*.

Andrew Halterman, Philip A Schrodt, Andreas Beger, Benjamin E Bagozzi, and Grace Scarborough. 2023. Creating custom event data without dictionaries: A bag-of-tricks. *International Studies Association Conference Paper*.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Reyyan Yeniterzi, Osman Mutlu, and Erdem Yörük. 2022. Challenges and applications of automated extraction of socio-political events from text (case 2022): Workshop and shared task report.

Suradej Intagorn, Anon Plangprasopchok, and Kristina Lerman. 2010. Harvesting geospatial knowledge from social metadata. In *ISCRAM*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Benjamin J. Radford. 2021. Regressing location on text for probabilistic geocoding. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 53–57, Online. Association for Computational Linguistics.

Clionadh Raleigh, rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of peace research*, 47(5):651–660.

Ralf Steinberger and Bruno Pouliquen. 2007. Cross-lingual named entity recognition. *Lingvisticæ Investigationes*, 30(1):135–162.

Hristo Tanev. 2022. Ontopopulis, a system for learning semantic classes. In *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 8–12.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems: 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008 London, UK, June 24-27, 2008 Proceedings 13*, pages 207–218. Springer.

Hristo Tanev, Vanni Zavarella, Jens Linge, Mijail Kabadjov, Jakub Piskorski, Martin Atkinson, and Ralf Steinberger. 2009. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *Linguamática*, 1(2):55–66.

Andreas Forø Tollefsen, Håvard Strand, and Halvard Buhaug. 2012. Prio-grid: A unified spatial data structure. *Journal of Peace Research*, 49(2):363–374.

Marc Wick and C Boutreux. 2011. Geonames. *GeoNames Geographical Database*.

Vanni Zavarella, Hristo Tanev, Ali Hürriyetoğlu, Peratham Wiriyathammabhum, and Bertrand De Longueville. 2022. Tracking COVID-19 protest events in the United States. shared task 2: Event database replication, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 209–216, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.