Mastering the Media Hype: Methods for Deduplication of Conflict Events from News Reports

Vanni Zavarella[1*], **Jakub Piskorski**[2], **Camelia Ignat**[1], **Hristo Tanev**[1] and **Martin Atkinson**[1]

[1]Joint Research Centre of the European Commission, Ispra, Italy
[2]Polish Academy of Sciences, Warsaw, Poland

Abstract

Machine coding of conflict event datasets has recently emerged as a time-effective method which can back up predictive models for conflict escalation at national and sub-national level. However, the event record duplication issue, caused by large news coverage of major conflict events, significantly degrades the accuracy of these datasets and makes them unreliable for micro-analysis of conflict processes. In this paper, we assess the effectiveness of two automatic approaches for mitigating the event duplication issue. The first approach (Cluster Linking) consists of linking news article clusters across time, prior to event extraction, while the second one (Event Linking) is based on classification and aggregation of related events. The comparative evaluation is performed by measuring the correlation of the output from an automatic event detection system with human-coded conflict events from the ACLED project, spatially aggregated on administrative units. We find out that, while both methods effectively reduce the automatic system's large outlier event and victim counts (with a slight prevalence of Event Linking), they can only increase the correlation coefficients with human-coded data significantly if coupled with an accurate and fine-grained geocoding module.

1 Introduction

The last decade has seen a surge of interest in models of socio-political violence and conflicts integrating the standard static indicators (e.g. census data) with more dynamic indicators such as time-stamped event records. This has stimulated the creation of several machine-coded event datasets, fully- or semi-automatically generated from news reports with relatively rich semantic representations (see [Leetaru and Schrodt, 2013], [Lorenzini *et al.*, 2016]). At the same time, several concerns have emerged towards the usability of machine-coded event datasets for micro-level modelling, particularly in the domain of political violence where spatial analysis has become standard. These range from inconsistency of data schemas, making datasets not directly comparable ([Wang *et al.*, 2016]), through source inconsistency over time, up to the more general problem of the validity of machine-coded data with respect to a Gold Standard of unique real-world events [Hammond and Weidmann, 2014].

In particular, a low correlation has been found with respect to human-coded reference data, upon quantitative analysis based on temporal-geographical aggregation (e.g. see [Hammond and Weidmann, 2014] for GDELT). This seems to be due to news-based event-coding systems being too sensitive to the intensity of media reporting and generating large sets of event duplicates or near-duplicates. In this paper we experiment with two automatic approaches for mitigating the event duplication issue. The first approach is based on linking clusters of news items, while the second one is based on classification and aggregation of related events. The comparative evaluation is performed by measuring the correlation of the output from an automatic event detection system with human-coded conflict events from the ACLED project [Raleigh *et al.*, 2010].

By making the extracted representation of complex, evolving processes such as conflicts closer to gold standards, these techniques contribute to the larger endeavour in the NLP community on automatic narrative extraction and construction from text [Jorge *et al.*, 2019].

The paper is structured as follows. Section 2 describes some background work on event dataset correlation studies and some existing approaches on tackling the event duplication problem. Section 3 briefly introduces an existing automatic event extraction engine and presents a correlation analysis vis-a-vis a Gold Standard dataset. Section 4 presents two approaches we deployed to increase the baseline correlation figures. In Section 5 we report on the impact of these methods on our target datasets. Finally, we end up with conclusions in Section 6.

2 Related work

Several publications recently focused on assessing the correlation of event datasets based on disaggregated event counts,

*Corresponding Author.

In: A. Jorge, R. Campos, A. Jatowt, A. Aizawa (eds.): Proceedings of the first AI4Narratives Workshop, Yokohama, Japan, January 2021, published at http://ceur-ws.org

for example [Ward *et al.*, 2013]. [Schrodt and Analytics, 2015] presents an extensive comparison of large well-known event datasets. [Hammond and Weidmann, 2014] apply spatio-temporal disaggregation of events incidents to assess whether GDELT data can approximate the spatial pattern of conflicts. We applied an adapted version of their correlation analysis in this paper.

[Schutte *et al.*,] address the usability and the presence of duplicates in various popular datasets of political and conflict events in the research community, such as ACLED, GDELT, and ICEWS[1]. A number of other papers refer to usability of the event conflict databases. For example, [Demarest and Langer, 2018] analyzed conflicts and social unrest in Africa, using event datasets.

The event duplication issue that we tackle here has been approached in Computational Linguistics by several works on co-reference resolution applied to event mentions [Lu and Ng, 2018]. Full-fledged event co-reference resolution is a harder semantic task than the one we deal with here and requires to take into consideration various deep linguistic features, due to the complexity of the event mentions. For example, [Lee *et al.*, 2012] resolve simultaneously noun phrase co-references and cross-document event co-reference, using an original algorithm which exploits clustering, pronoun resolution and semantic roles labeling. An unsupervised graph-based method for event co-reference is presented in [Bejan and Harabagiu, 2010]. [Zhang *et al.*, 2015] use both textual and visual scene similarity features, when resolving co-reference news captions. Another interesting method for linking similar events based on machine learning and various similarity features has been presented in [Piskorski *et al.*, 2018]. We deploy a modified version of this approach for our Event Linking in Section 4.

3 Correlation across datasets

For our correlation analysis we focus on two major violent conflicts that recently plagued the African region: the Libyan Civil War and the Mali War. For each of them, we compare the datasets generated by a fully automatic event detection engine with Gold Standard data coded by human experts within the ACLED project [Raleigh *et al.*, 2010]. For our experiments we use the output of the English language instance of NEXUS [Atkinson *et al.*, 2017], a Joint Research Center in-house multilingual system that has been running continuously since 2007. NEXUS is a finite-state rule based event extraction engine that processes in real-time the title and lead sentences of monolingual clusters of news articles (for up to 10 languages) and outputs an event description template corresponding to the main event reported in each cluster. The clusters are computed every 10 minutes on a 4 hour window of RSS feeds (title and lead sentences) of news sources by the Europe Media Monitor (EMM), a large-scale multilingual news aggregation engine that gathers articles from ca. 7000 sources (from local to global level) in 60 languages on a 24/7 basis [Atkinson and Van der Goot, 2009]. Event templates include two main slots: *Event_Type* and *Event_Location*, together with other event-type specific descriptive and numerical slots such as *Number_Dead,Number_Injured* etc. NEXUS features a rule-based event geocoding algorithm that integrates (a subset of) the Geonames gazetteer[2] for place name matching and a number of article-level disambiguation heuristics for geo-disambiguation.

Because NEXUS does not feature the same data schema and event taxonomy as ACLED, some type mapping was performed. Table 1 describes the datasets and the type filtering that generated them. Moreover, while both NEXUS and ACLED encode geographical information as a hierarchy of administrative level components, like in the following example:

Populated_Place=Kidal
Admin1=Tessalit
Admin2=AmiAdjelhoc
Country=Mali

the components are not id-indexed. Therefore we normalized geographical references by matching the name variants on the high coverage Geonames gazetteer [3].

The machine-coded event datasets (including instructions on how to access the underlying news stories from which they were extracted) can be accessed at: http://labs.emm4u.eu/events.html

In order to set a baseline correlation between Nexus and the ACLED data, we aggregate event counts per time/space cells, where time is either a week or a month range, and the space is either a Province or Region level administrative unit. Figure 1a and 1b below visualize the dynamics of the Libyan and Malian conflict escalation/de-escalation by showing on each week (month) the total number of province and regions, respectively, experiencing one or more violent event incidents.

While rather standard, this analysis is highly affected by the relatively more coarse-grained event geocoding of NEXUS compared to the human-coded Gold Standard [4]. Moreover, it does not consider variance in conflict intensity estimation within time/space units, which is crucial for micro-level analysis of conflicts at sub-national level. Conflict intensity can be measured by absolute event counts and by total victim counts. Therefore, in Figure 2a and Figure 2b we plot total weekly event counts and victim counts (respectively) of NEXUS compared to statistics from ACLED data. The same figures are reported, on a monthly base, in Figure 3a and Figure 3b for Mali War.

Table 2 reports a number of error rate measures and correlation coefficients between ACLED and NEXUS datasets for the two target conflicts.

Overall, the analysis shows a moderate to strong level of correlation for event counts and a correlation from negligible

[1] See http://www.lockheedmartin.com/us/products/W-ICEWS/W-ICEWS_overview.html for more information.

[2] http://geonames.org/.

[3] We filtered out a total of 5% of ACLED events that could not be normalized with respect of the Geonames resource.

[4] The range of geographical distribution of NEXUS events is much lower because whenever it fails to locate an event at the exact populated place level, it backs off to the capital city of the most specific administrative unit it could detect.

CONFLICT	START DATE	END DATE	ACLED Events *Air/drone strike, Armed clash, Protest with intervention, Bombing, Mob violence,Sexual violence, Suicide bomb, Violent demonstration, Excessive force against protesters, Shelling/artillery/missile attack,Grenade, Attack,Remote explosive/landmine/IED Chemical weapon*	Nexus Events *Assassination,Terrorist Attack, Suicide Attack,Air/Missile Attack Bio Chemical Attack,Heavy Weapons Fire, Shooting,Execution, Armed Conflict,Firebombing*	Nexus_CL	Nexus_RLT
Libyan Civil War	17-02-2011	23-11-2011	466	802	342	527
Mali War	16-01-2012	15-04-2015	525	496	-	455

Table 1: The conflict event datasets and mapped event types from ACLED and NEXUS.

to weak for victim counts, according to standard interpretation of correlation coefficients [Hinkle *et al.*, 2003].

By looking at the curves, one can notice that the automatic system is less sensitive to low conflict signals. This might be due to both a general recall deficit of rule-based approaches and a granularity deficit of the underlying geocoding algorithm. Minor conflict incidents receive relatively lower media reporting and small sized news clusters are more likely to produce false negatives from a low-recall automatic system. Moreover, even in cases where such small signals are detected, there is a significant chance that the system lacks the geographical knowledge to correctly locate them. On the other hand, NEXUS tends to over-generate events at high signal intensity points, possibly because it is unable to normalize the increased stream of media reporting coverage on major event incidents. This, combined with the invalid extraction of outlier figures, makes the error rates particularly poor for victim counts. These two drawbacks combined highly hinder the usability of NEXUS-generated datasets for quantitative conflict modelling at sub-national level.

4 Deduplication methods

We now experiment with combining Nexus with two alternative methods, in order to mitigate the issues underlined by the correlation analysis. The first one, Cluster Linking, is a pre-processing step that consists of further aggregating over a given time window the daily news clusters on which NEXUS is run. The rationale is that major event incidents generate complex news stories spanning over several days and following a range of related topic threads: being able to track *a priori* these stories spares the downstream event extraction engine the burden of detecting and aggregating event co-references.

The second method, Event Linking, is a post-processing step that approximates event co-reference resolution by deploying a classifier for event Relatedness, and then clusters the resulting graph, aggregating the output event classes.

4.1 Cluster linking

The cluster linking is part of a larger multilingual application that identifies equivalent news clusters across languages and over time. It uses language-independent features as weighted vectors, and calculates the news cluster similarity as their linear combination. In this experiment we only apply the historical cluster links on English news.

The cluster representation is based on the following features and their weights: (1) Named Entities: person names and organisations; (2) Geolocations: geographical places detected in each cluster; (3) Content categories: predefined thematic categories that are assigned to each cluster; (4) Automatically produced translations into English for cross-lingual linking, or word tokens for monolingual linking, based on the titles and the short descriptions of each article in a cluster; (5) Eurovoc categorisation: the whole clusters are automatically indexed with Eurovoc categories, using the freely available software JEX[5]. (6) Combined feature: Name Entities + Geolocations. We have explored and evaluated different weighting methods: (a) normalized frequency; (b) log-likelihood, that compares the term frequency in the cluster with the frequency in a reference corpus and (c) TF-IDF, that is proportional to the term frequency quantified by the inverse function of the number of documents in which the term occurs. A different weighting technique is selected for each feature: log-likelihood (using a reference corpus) for feature (2) and (4) and tf-idf for (1),(2) and (6). For each of these features separately, the similarity is computed between two clusters as the cosine between the weighted vectors generated by the feature modified by a penalty metrics:

$$Sim_{FEATURE}(x, y) = cosine(x, y) * penalty(x, y)$$

The penalty metrics are: (1) Dimension Penalty, that decreases the similarity value in case of low dimensionality of the feature vectors (low numbers of entities found), (2) Jaccard Penalty, based on Jaccard index - that considers the ratio between the number of common terms and the total number of terms, (3) combination of 1 and 2 as the product, the minimum or the maximum of the values. After optimising each individual feature vector similarity, the global similarity between two clusters is calculated as a linear combination of the individual feature vector similarities:

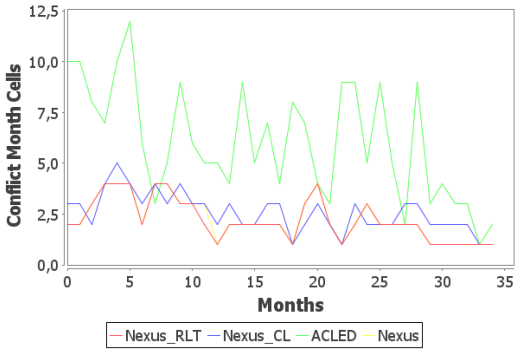
$$SimCluster = w_1 * SimNE + w_2 * SimGL + w_3 * SimCat + w_4 * SimTransl + w_5 * SimEvcDesc + w_6 * SimNG$$

Further, for a given period of time, a graph of clusters is generated by selecting the ones with the similarity above a given threshold and the graph is expanded to its transitive closure. Each graph represents a group of clusters that are similar.

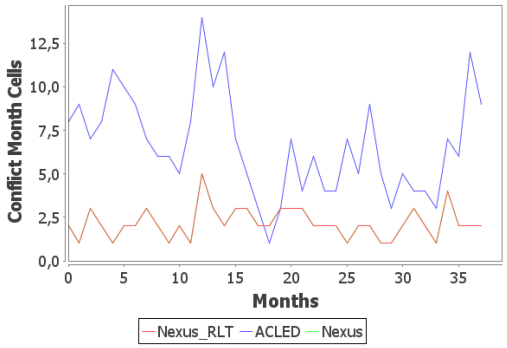
4.2 Event Linking

For the sake of computing pairs of related events we used a Random Forest-based classifier trained on a corpus of ca. 23K event pairs tagged as related or unrelated, where the

[5]https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer.



(a) Weekly province conflict unit counts for Libyan Civil War.

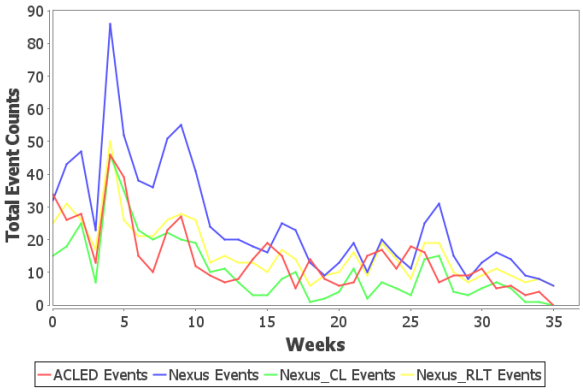


(b) Monthly province conflict unit counts for Mali War. Nexus and Nexus_RLT are completely overlapping in this case.

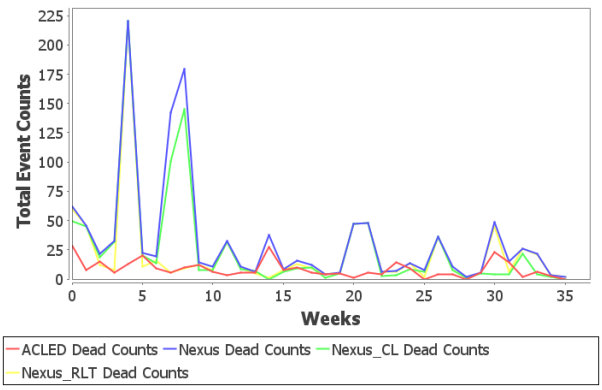
Figure 1

		Nexus				Nexus_CL				Nexus_RLT			
		RMSE_Prov	RMSE_Reg	r	ρ	RMSE_Prov	RMSE_Reg	r	ρ	RMSE_Prov	RMSE_Reg	r	ρ
Libya Civil War	Events	8.556	9.784	0.799	0.628	3.956	4.827	0.767	0.523	4.458	5.390	0.797	0.622
	Dead	8.297	8.365	0.048	0.445	8.162	8.229	0.029	0.15	8.053	8.121	0.083	0.304
Mali War	Events	4.955	5.010	0.64	0.396	n/a	n/a	n/a	n/a	4.545	4.691	0.604	0.349
	Dead	4.268	4.443	0.191	0.162	n/a	n/a	n/a	n/a	4.225	4.40	0.162	0.158

Table 2: Error rates and correlation coefficients of event and victim counts for Libya and Mali conflicts, at different level of geographical aggregation. *RMSE_Prov* is the Root Mean Squared Error computed in time-Province units, *RMSE_Reg* is the Root Mean Squared Error computed in time-Region units, *r* is Pearson correlation coefficient, while ρ is Spearman's rank correlation coefficient. The log values of Root Mean Squared Error are shown for Dead counts.



(a) Total event counts, per week units for Libyan civil war.



(b) Total victim counts, per week units for Libyan civil war.

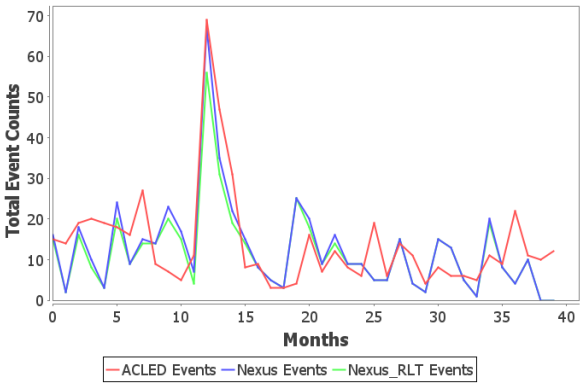
Figure 2

events are represented as texts consisting of the title and 1-2 lead sentences from news articles reporting on crisis and security-related events. In particular, for training the classifier a set of about 15 features was exploited including, i.a., string distance metrics (e.g., Longest Common Substrings), features that exploit knowledge bases (e.g.,WordNet, BabelNet) to compute, e.g., WORDNET-word overlap, Named-Entity overlap, Hypernym overlap, and some corpus-based event similarity metrics, e.g., Weighted Word overlap, which measures the overlap of words between the two texts, where words bearing more content (i.e., appear more frequently in the domain corpus) are assigned higher weight. The trained classifier obtained 91.5% f measure on hold-out test data consisting of approximately 20% of the entire event corpus. Further details on the classifier can be found in [Piskorski *et al.*, 2018].

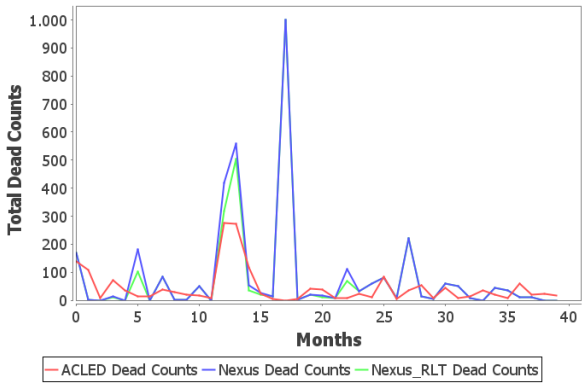
5 Evaluation

5.1 Cluster Linking application

In our experiment, the cluster linking module was applied to all news clusters geo-coded in Libya from the period 02/2011-11/2011. Only four (out of six) features were con-



(a) Total monthly event counts for Mali war.



(b) Total monthly victim counts for Mali war.

Figure 3

sidered: named entities, geolocations, tokens and the combination of named entities and geolocations. The usage of the other two features (categories and Eurovoc descriptors) are recommended in cross-lingual experiments with poor or no translation resources. Otherwise, in the monolingual case, they are not increasing significantly the cluster linking precision (and Eurovoc categorisation is a heavy computation task). We have used two thresholds: one as a cut-off for the clusters in the same day (monolingual daily threshold 0.72) and the second for the historical linking, to select similar cluster over time (historical linking threshold 0.62). Firstly, we have selected all the pairs of clusters with a similarity above the thresholds and then the new cluster links were added by transitivity. Groups of clusters have been generated considering all similar clusters. NEXUS was then run on these expanded clusters and one main event was extracted from each. This produced a lower size dataset for the Libyan conflict, as it can be seen in Column *Nexus_CL* of Table1).

Event Linking application The pairs of related events returned by the classifier were transformed into an undirected graph and path distance was used as similarity metrics for applying Agglomerative Clustering (with cluster cardinality set to 50 for both datasets[6]). We finally applied this clustering to time-based partitions of the datasets comprising events within 3 days intervals and geocoded in the same region and merged the resulting aggregated events. The final datasets are referred to as *Nexus_RLT* in Table1).

As it can be visually seen in Figures 2a through 3b, both methods seems to get the curves for cumulative weekly and monthly event and victim counts closer to the ACLED data, for both conflicts. This is particularly true with respect to event counts, while victim counts seem to suffer from some outlier values extracted by NEXUS. Root Mean Squared Error figures in Table 2 confirm that the application of both deduplication techniques produces a systematic reduction of the absolute error rate of NEXUS, more significant for event than for victim counts.

[6]We used https://networkx.github.io/ and https://scikit-learn.org/ libraries for graph modelling and clustering implementation, respectively.

Event Linking seems to be more robust to outliers as these can be mitigated by slot value merging heuristics, applied downstream of the extraction engine. This explains why *Nexus_RLT* almost doubles the correlation coefficients for Libya victim counts with respect to *Nexus*, as outliers are more likely to be extracted for victims. For event counts instead, Cluster Linking produces the higher drop in error rates.

On the other hand, none of the methods is able to consistently increase the correlation coefficients *Nexus* with ACLED data by a significant factor. We hypothesize this might be due to the inaccuracy of the event location information. As we mentioned in Section 3, the range of province-level spread of NEXUS events is only 10% and 20% of the ACLED data for Libyan and Mali conflicts, respectively. In this scenario, the information aggregation achieved by either Cluster Linking or Event Linking is applied at a too coarse-grained geographical level and might actually over-compress the signal, by underestimating the total event counts.

6 Conclusions

The usability of news-based, automatic coding of event datasets for conflict analysis at sub-national level has been questioned in the conflict analysis research community.

We have shown how the application of two linguistically light-weight text processing modules can mitigate, although only partially, some of the standard flaws of news-based event coding engines, namely the over-generation of event duplicates at peaks of news reporting intensity.

The cluster linking method is unsupervised and the only customization to the event duplicate detection task consisted of setting up plausible similarity thresholds, with no optimization performed.

The event linking is based on a trained supervised classifier, however the training set was not overlapping with the two target conflict datasets, which means that it shared virtually no semantic context (e.g. named entities) with it.

Therefore, we estimate that both approaches can well generalize and be deployed to mitigate event duplication across different datasets.

On the other hand, customizing the presented methods so

as to optimize the boost in correlation coefficients of target datasets is a promising direction to explore. For instance, we plan to sample the NEXUS event records at peaks of generation (with respect to Gold Standard) in order to collect annotated data for training a more specialized event duplicate classifier. In this respect, we expect semantic features such as geolocation and time stamp might to turn out being highly discriminative.

Overall, we estimate that the effectiveness of the presented methods can be better assessed when coupled with an event extraction engine with an underlying high-accuracy, fine-grained geocoding module. Therefore we plan to re-run the correlation benchmark after moderating the NEXUS event location slots.

References

[Atkinson and Van der Goot, 2009] Martin Atkinson and Erik Van der Goot. Near real time information mining in multilingual news. In *Proceedings of the 18th international conference on World wide web*, pages 1153–1154, 2009.

[Atkinson *et al.*, 2017] Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65, 2017.

[Bejan and Harabagiu, 2010] Cosmin Adrian Bejan and Sanda Harabagiu. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, 2010.

[Demarest and Langer, 2018] Leila Demarest and Arnim Langer. The study of violence and social unrest in africa: A comparative analysis of three conflict event datasets. *African Affairs*, 117(467):310–325, 2018.

[Hammond and Weidmann, 2014] Jesse Hammond and Nils B Weidmann. Using machine-coded event data for the micro-level study of political violence. *Research & Politics*, 1(2):2053168014539924, 2014.

[Hinkle *et al.*, 2003] Dennis E Hinkle, William Wiersma, and Stephen G Jurs. Applied statistics for the behavioral sciences (vol. 663), 2003.

[Jorge *et al.*, 2019] Alípio M Jorge, Ricardo Campos, Adam Jatowt, and Sérgio Nunes. Information processing & management journal special issue on narrative extraction from texts (text2story), 2019.

[Lee *et al.*, 2012] Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics, 2012.

[Leetaru and Schrodt, 2013] Kalev Leetaru and Philip A Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer, 2013.

[Lorenzini *et al.*, 2016] Jasmine Lorenzini, Peter Makarov, Hanspeter Kriesi, and Bruno Wueest. Towards a Dataset of Automatically Coded Protest Events from English-language Newswire Documents. In *Paper presented at the Amsterdam Text Analysis Conference*, 2016.

[Lu and Ng, 2018] Jing Lu and Vincent Ng. Event coreference resolution: A survey of two decades of research. In *IJCAI*, pages 5479–5486, 2018.

[Piskorski *et al.*, 2018] Jakub Piskorski, Fredi Šarić, Vanni Zavarella, and Martin Atkinson. On training classifiers for linking event templates. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 68–78, 2018.

[Raleigh *et al.*, 2010] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660, 2010.

[Schrodt and Analytics, 2015] Philip A Schrodt and Parus Analytics. Comparing methods for generating large scale political event data sets. In *Text as Data meetings, New York University, 16–17, 2015*, pages 1–32, 2015.

[Schutte *et al.*,] Sebastian Schutte, Howard Liu, and Michael D Ward. The more the merrier? addressing duplications in automated event data.

[Wang *et al.*, 2016] Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503, 2016.

[Ward *et al.*, 2013] Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. Comparing gdelt and icews event data. *Event Data Analysis*, 21(1):267–297, 2013.

[Zhang *et al.*, 2015] Tongtao Zhang, Hongzhi Li, Heng Ji, and Shih-Fu Chang. Cross-document event coreference resolution based on cross-media features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 201–206, 2015.