Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 2, 2024
A Regional Conference of the Society for Industrial and Systems Engineering

An Output-Based Evaluation Framework for Event Extraction Models

Seth Brown, Daniel Hoffmann, Ian Mickle, Gunnar Schwab, and Ian Kloo

Department of Systems Engineering,
United States Military Academy,
West Point, NY 10996

Corresponding author's Email: mickle.ian@outlook.com

Author Note: This work was completed as part of a USMA capstone project in the Department of Systems Engineering. We would like to thank the department and our client for giving us this opportunity to research this important field. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: Event extraction models aid in analysis and decision-making in many domains; however, there are few established evaluation techniques or frameworks that compare their potential utility. We demonstrate a methodology for determining which of two event models is more useful for enhancing situational awareness using public news data. Specifically, we evaluate the Global Database of Events, Langauge, and Tone (GDELT) model against a proprietary model developed internally by a federal agency. This project looks beyond academic benchmarks and directly evaluates the outputs of the two models regarding potential analytic utility. The GDELT model is more immediately usable due to its greater specificity and stronger entity resolution capabilities for actors and events. Beyond assessing these specific models, the evaluation methodology described in this paper could be useful for evaluating other large, opaque natural language processing models in terms of their usability in real-world applications instead of measuring against academic benchmarks.

Keywords: Natural Language Processing, Event Model, Transformer Model

1. Introduction

With an increased focus on data-based decision-making throughout the United States Army and DoD, natural language processing (NLP) tools continue to gain popularity. The NLP field is comprised of theoretical and computational techniques to enable a computer to analyze human languages (Liddy, 2001). In this paper, we will be comparing a specific type of NLP model that seeks to extract events and actors from text data (S. Yang, Feng, Qiao, Kan, & Li, 2019). These models are useful for parsing large text corpora and can be used to generate situational awareness for a specific location and time.

Elements of the U.S. government currently employ an event extraction model called Global Database of Events, Language, and Tone (GDELT), created by researchers at Georgetown University (GDELT Project, n.d.), to parse events from news data. Recently, another U.S. agency developed a closed-source model (identified as Closed Source Event Model or CSEM in this paper) with the same intent. Both models are performant on NLP benchmarks (i.e., they accurately extract events and actors from human-tagged data), so this study will instead focus on how well analysts could use the outputs from each model to create global situational awareness.

The remainder of this paper will describe some background on key NLP techniques used by GDELT and CSEM, provide the context for a case study used to evaluate the model outputs, explain the comparison methodology used in this work, present our results, and discuss how our findings can inform model selection and future evaluations of similar models.

2. Background

Before proceeding with our study, it is important to provide background on some key aspects of NLP as well as the GDELT and CSEM models. This section also provides a brief contextual background on an event we will use as a case study: the attack on the Al-Shifa hospital in Gaza.

2.1. Natural Language Processing: Event Extraction

NLP is a rapidly developing field in machine learning that leverages human language as inputs and outputs (Liddy, 2001). Common NLP tasks are to ”paraphrase an input text [...] translate the text into another language [... and] answer questions from the contents of the text” (Liddy, 2001).

This paper will focus on event extraction, which takes natural text as input and outputs a list of events and actors contained in that text. Event extraction relies on span extraction techniques, which represent text as ordered tokens and classify groups of tokens as events or actors. This type of model is also able to identify relationships within and without a sentence. This is done with contextualized embedding using transformer models like BERT (S. Yang et al., 2019). These models first encode the text by processing a sentence and the neighborhood of surrounding sentences. Next, span enumeration concatenates the tokens' endpoints, effectively finding a smaller chunk of text within a larger input. A span graph propagation structure, a graphic where each x variable has a range of y values, is then used to dynamically present the model's best guess of the relationships within the chunk. Lastly, the model conducts multi-task classification by feeding the representations as input into scoring functions, in this case a two layer feed forward neural network, outputting the relevant pair of embeddings (Wadden, Wennberg, Luan, & Hajishirzi, 2019).

2.2. GDELT

GDELT is an event model used to update a publicly available database daily from millions of news articles. The database contains historical data from as far back as 1979 and can be easily separated by year, location, or event (GDELT Project, n.d.). The longstanding public availability of GDELT has made it the gold standard for global news synthesis; however, the model architecture of GDELT is opaque. Specifically, it is unknown if GDELT has been modified to include recently developed NLP techniques or if it uses a legacy framework.

2.3. CSEM

CSEM is a closed-source event model developed internally by a government agency in conjunction with outside academic experts. It uses a mixture of transformer models to perform event extraction, but the specific methodologies used to train these transformers are unknown.

2.4. Al-Shifa Case Study

The methodology described in the next section leverages a case study to compare the outputs of GDELT and CSEM. This section will provide a brief background on the events surrounding our case study. The Al-Shifa Hospital is a civilian hospital located in Gaza City. It is the largest hospital in Gaza and is one of the few remaining semi-operational hospitals in the area. Following the Hamas attacks on Israeli civilians in October 2023, there were reports that Hamas militants were staging within and beneath the hospital. To clear the Hamas militants, Israel conducted a raid in November 2023. The raid, compounded with indirect fires, led to casualties of Hamas combatants and Palestinian civilians (Reuters, 2024).

The Al-Shifa Case study was chosen to help facilitate comparison to real-life events. The Al-Shifa Hospital was only referenced in 18 of the 10,000 article data pool which enables the team to directly compare the articles to the model outputs.

3. Methodology

This section describes our data collection/processing strategy, a methodology for comparing model ontologies, and a case-study approach for output comparison.

3.1. Data Collection and Processing

The goal of this research is to evaluate GDELT and CSEM by comparing their outputs. To facilitate direct comparison, it was important to process the same news articles with each model. The study team did not have access to the GDELT model code but could access the public GDELT database populated with output daily for a selection of news articles. We started by filtering the GDELT data to contain articles in the weeks following the Hamas attack on Israel (October 7, 2023). Next, we selected only events that contained either Israel, Hamas, Gaza, or Palestine (and any related terms). Finally, we took only the first 10,000 matches to keep the data at a reasonable size.

GDELT does not provide the full news articles used to extract events, but the database does contain the URL to the original source. Using these URLs, we used a Python-based web scraping approach to collect the full text of each article. We obtained the full codebase for the CSEM model and ran the selection of 10,000 articles through the model. This process left the team with input (news articles) and output (events and actors) pairs for each of the 10,000 articles for GDELT and CSEM.

For the Al-Shifa case study, we further subset our data to 18 news articles that directly discussed the Israeli raid or Hamas presence in the hospital.

3.2. Ontological Comparison: Event Comparison

Both GDELT and CSEM label and extract events, but they use different ontologies to name these events. GDELT categorizes events into 20 major buckets (e.g., Make Public Statement, Appeal, Assault, etc.). These are further stratified into sub-events (e.g., Appeal for Material Cooperation). CSEM's ontology only has a single level at a similar granularity to GDELT's sub-events. After processing 10,000 news articles with each model, we found 202 unique GDELT event types and 104 unique CSEM event types.

There are no event types in either model with the exact same name, just similar meanings, making the comparison difficult and time-consuming. The first step in our comparison was a one-way translation of event types using human evaluators. For every event type in CSEM, we looked through the entire list of GDELT event types for similar events.

The second and final step of the comparison used a combination of GPT-4 and human verification. Using the same premise as the initial human evaluation, we prompted GPT-4 to compare each event type to the full list of event types on the other model. This was conducted bidirectionally, with each run generating two lists: one list of every event in CSEM with all equivalent events in GDELT and vice versa. These lists were joined and deduplicated. Finally, two human evaluators independently reviewed every proposed equivalence from the compiled list and tagged them as correct or incorrect. The labels where both humans agreed with the model were considered a true match.

3.3. Output Comparison

To compare the outputs generated by each model, we developed a methodology that explores the extracted events and actors through multiple lenses. Applying this methodology to the outputs of both models allowed us to compare and contrast the models against each other. Furthermore, because we applied both models to a subset of news articles pertaining to a known event (the Al-Shifa hospital attack), we were able to compare the model results to ground truth, ultimately showing which model has more analytic value.

We began our output processing by simply looking at the frequency of events and actors. Next, we created bipartite knowledge graphs to link events to actors. These allowed us to move beyond simply counting the extracted events to study the interactions that the actors and events had according to each model using dynamic network analysis techniques that characterize network structure in terms of density and centrality at both the node and network levels. Finally, we used network analytic methods to fold the bipartite networks to create actor-to-actor and event-to-event networks. These networks allow for more advanced network analytics, including hierarchical clustering and Louvain community detection to group similar actors and events (Tian, Zhang, Fei, Song, & Feng, 2021).

4. Results

4.1. Ontological Comparison: Event Type Validation

In the first step of our ontological comparison, we compared each CSEM event type to the full list of GDELT event types. Overall, 68% of GDELT events had at least one similar event type in CSEM. The major event types that we expected to see (e.g., Military Attack, Diplomatic Cooperation, etc.) were present in both models.

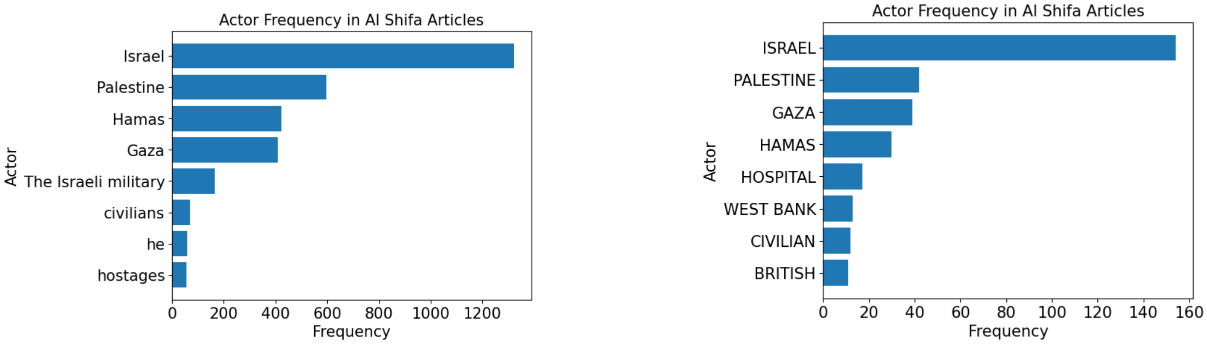
To expand our ontological comparison and make a comprehensive two-way comparison of all event types, we leveraged GPT-4. This model found 1,194 matches the CSEM and GDELT event types. Two human raters then validated every comparison in the GPT-4 output, reaching a consensus in 81% of cases. The remaining 19% were judged by the human raters to be incorrect labels and were discarded.

4.2. Output Comparison: Al-Shifa Case Study

The following sections show analytic outputs that we would expect an analyst to find useful when using an event extraction model to gain situational awareness from news data. All of these plots were constructed using the data from 18 news articles in our dataset that directly discussed the Al-Shifa hospital raid.

4.2.1. Frequency Analysis

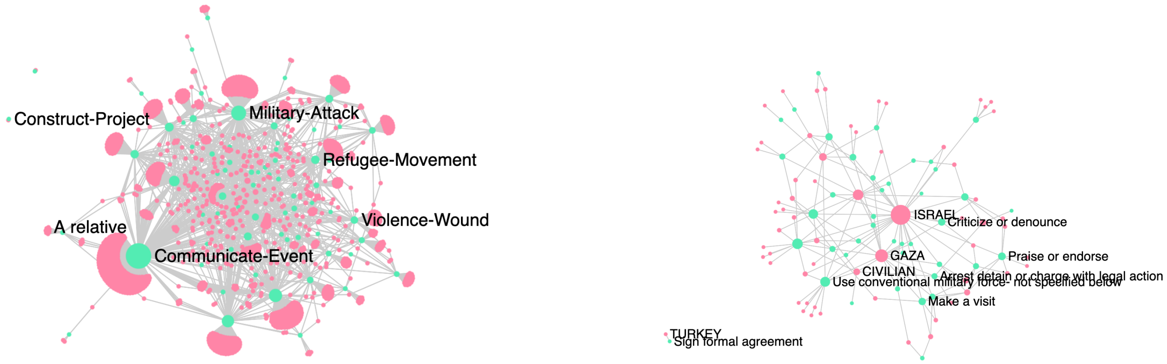
Figure 1 shows a bar graph depicting the frequency of actor identification in the Al-Shifa information space for both CSEM and GDELT. We observed significant overlap in the key actors from each model using relative frequency but noted that the number of actors was much higher for the CSEM model. The CSEM model also provided many similarly named actors (e.g., Israel and Israeli), while GDELT resolved the actors to single, common names. Finally, the CSEM output contained many pronouns as actors, reducing the functionality of the model.



(a) CSEM Actor Bar Graph

(b) GDELT Actor Bar Graph

Figure 1: Actor Frequency Bar Graphs



(a) CSEM actors to events network

(b) GDELT actors to events network

Figure 2: Actor to Event Networks for CSEM and GDELT.

4.2.2. Network Diagrams

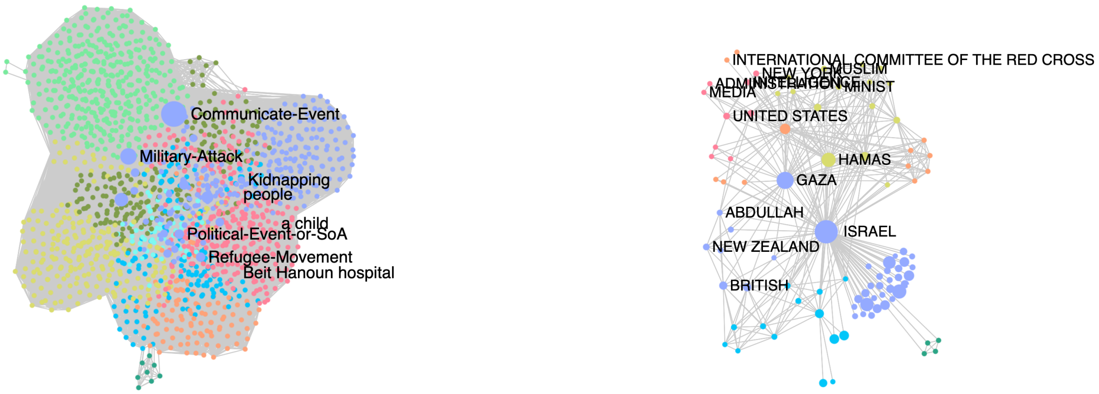
Figure 2 shows two networks for each model: a bipartite actor-to-event network and an actor-to-actor network that resulted from folding the initial bipartite network. In both networks, red nodes are actors, and green nodes are events. In Figure 3, color corresponds to groups assigned using the Louvain clustering algorithm. In all networks, node size corresponds to degree centrality (the count of edges associated with the node). An edge in the networks in Figure 2 means indicates the model linked an actor to an event. The edges in the networks in Figure 3 mean a pair of actors were linked to the same event.

Figures 2a and 2b show that the CSEM contains many more actors (1,796) and events (77) than GDELT (63 actors and 44 events). However, the CSEM has a much lower density coefficient (0.00157) than GDELT (0.0315). This coefficient measures the ratio of existing edges to the number of edges that would exist in a fully connected network. The CSEM network (characterized by large green event nodes surrounded by innumerable disconnected pink actors) shows that events are more central than actors, while the opposite is true for GDELT.

Similarly, Figures 3a and 3b show that CSEM identified many more actors than GDELT. The density coefficient of the CSEM network was 0.125 compared to 0.078 for the GDELT network, suggesting that the actors in the CSEM output tended to be more related to each other through the same event types. The Louvain algorithm identified seven sub-communities in the CSEM network and five in the GDELT network.

5. Discussion

The ontological comparison shows a strong overlap between the types of events that each model can identify, and we did not find any major event types to be missing from either model. However, while we found that the same basic event types were present, we also noted a significant difference in the scope of the events for each model. GDELT focuses on conflict and relationships between states, while CSEM has more domestic events. Additionally, CSEM contains a large number of general



(a) CSEM actor to actor network

(b) GDELT actor to actor network

Figure 3: Actor to actor networks for CSEM and GDELT.

events, such as ’Communicate Event’ which often appears in the data. The GDELT events were much more specific. Overall, the ontological comparison favors GDELT because the event types are more specific, and they capture all of the same major topics as CSEM. The CSEM event types seem too general, and we question their ability to support analysis. CSEM, however, would be preferred in cases where there is a domestic, non-conflict focus.

The network comparison further highlights the problems posed by CSEM’s focus on high-level events. Figure 2a shows that the largest node in the network is Communicate-Event. Recall that these networks were built from 18 news articles surrounding the Al-Shifa hospital attacks, and none of the articles contained what we would consider a major communication event. It seems that the model associates any person who may have said something in any context with a communication event. In contrast, it is very clear from looking at the GDELT network (Figure 2b) that the news articles are about a military/diplomatic event involving Israel and Hamas and that there is civilian involvement.

Similarly, the actor-to-actor network generated from the GDELT data was much more useful than the one generated by CSEM (Figure 3). Both networks show ”hospital” as a central node, but the CSEM network contains so many connected nodes that it is not useful for understanding the specific event. This problem is caused by the fact that CSEM associates so many actors with ”communicate event” that folding creates a network where almost all actors share a relationship. In effect, CSEM over-saturates results by tagging actors in excess to general events, which makes the outputs difficult to use. The folded GDELT network also shows the major actors connected to the hospital but displays a much clearer picture of the major actors involved. In conjunction with the network in Figure 2, the GDELT networks would be useful in gaining overall situational awareness of the Al-Shifa raid.

In general, CSEM identified many more actors and events than GDELT, but many of these failed to add insight. For example, the model tags pronouns (e.g., he, she, etc.) as actors but does not resolve that context to the actual actor the pronoun is referring to. Furthermore, the CSEM model does not attempt to resolve entities, leaving the analyst to deduplicate things like ”Israel” with ”Israeli,” which is not a trivial task. It is possible that with significant post-processing, we might be able to extract useful information from the CSEM output. Still, it is unclear how we could resolve the event over-tagging issue that results in overly dense, uninformative actor-to-actor networks.

6. Limitations and Future Work

While this study was able to develop a useful methodology for evaluating event models, we were largely constrained by a lack of access to the GDELT code base. Without the code, we were forced into a complicated process to find and scrape all of the documents used as inputs for GDELT using the methodology described in Section 3.1. This took considerable time and effort and ultimately limited the number of case studies we were able to perform. Future studies should look to gain access to the GDELT source code and run additional case studies. Focusing on a single case study limits the generalizability of our results; however, we cannot see any reason to believe that our findings would be unique to the Al-Shifa case study.

Another limitation of this study was limited access to the analysts who use public news data in their current workflows. The analytic products we generated are our best attempt at generating products that we think would be useful, but we were unable to validate them against real analyst workflows. As we continue, we hope to develop a more robust common analytic methodology that could be used to perform case study evaluations of model outputs. The work shown in this paper should be considered a motivating proof of concept for future work in this space.

Finally, due to time limitations, we were unable to dedicate a significant amount of time to writing post-processing scripts for either model. It is possible that the apparent weaknesses of CSEM could be resolved through a major post-processing effort that we did not have time to perform. We believe that this is something the model development team should address instead of leaving it for downstream analysts and users. Still, we acknowledge that there may be value in the model outputs that we were unable to identify due to this time constraint.

7. Conclusion

This paper shows how ontological comparisons and case studies performed on model outputs can be useful in determining a model's overall utility. Both models evaluated in this paper performed well on labeled academic datasets, but it is clear that the GDELT model is superior for generating analytics to provide situational awareness. With the growth of closed-source models (e.g., OpenAI's GPT models) and open-source models that are so complex that their architectures are difficult to understand, output-based comparisons (like the methodology demonstrated in this paper) present a useful way to evaluate these models in terms of their usability in specific analytic workflows. Future work should build upon the work presented in this paper to create robust, analyst-validated workflows that could be used to perform future evaluations or the application of an output-cleaning model to make CSEM more usable.

8. References

Diaz, Jaclyn. (2024). *Israel's military launched an overnight raid on gaza's largest hospital.* Retrieved from https://www.wpr.org/news/israels-military-launched-an-overnight-raid-on-gazas-largest-hospital

GDELT Project. (n.d.). *About gdelt.* Retrieved from https://www.gdeltproject.org/about.html

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of library and information science* (2nd ed.). New York: Marcel Decker, Inc.

Liu, J., Chen, Y., Liu, K., Bi, W., & Liu, X. (2020, November). Event extraction as machine reading comprehension. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 1641–1651). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.128 doi: 10.18653/v1/2020.emnlp-main.128

Reuters. (2024). *Israeli military says it killed 90 gunmen at Gaza's al-Shifa hospital.* Retrieved from https://www.reuters.com/world/middle-east/israeli-military-says-it-killed-90-gunmen-gazas-al-shifa-hospital-2024-03-20

Robinson, Kali. (2024). *What is hamas?* Retrieved from https://www.cfr.org/backgrounder/what-hamas

Schwarz, Birgit. (2024). *Interview: Building the evidence for crimes committed in israel on october 7.* Retrieved from https://www.hrw.org/news/2024/01/31/interview-building-evidence-crimes-committed-israel-october-7

Tian, X., Zhang, J., Fei, J., Song, X., & Feng, J. (2021). An improved louvain algorithm for community detection. *Mathematical Problems in Engineering*, *2021*, 1485592. Retrieved from https://doi.org/10.1155/2021/1485592 doi: 10.1155/2021/1485592

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc

Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). *Entity, relation, and event extraction with contextualized span representations.*

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C.-C. J. (2019). Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, *8*, e19. doi: 10.1017/ATSIP.2019.12

Yang, P., Cong, X., & Liu, Z. S. F. X. (2021). *Enhanced language representation with label knowledge for span extraction.*

Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019, July). Exploring pre-trained language models for event extraction and generation. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5284–5294). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1522 doi: 10.18653/v1/P19-1522