Reliability of Methods for Extracting Collaboration Networks from Crisis-related Situational Reports and Tweets

Ly Dinh
School of Information
University of South Florida
lydinh@usf.edu

Sumeet Kulkarni
School of Information Sciences
University of Illinois at Urbana-Champaign
sumeetk2@illinois.edu

Pingjing Yang
School of Information Sciences
University of Illinois at Urbana-Champaign
py2@illinois.edu

Jana Diesner
School of Information Sciences
University of Illinois at Urbana-Champaign
jdiesner@illinois.edu

ABSTRACT

Assessing the effectiveness of crisis response is key to improving preparedness and adapting policies. One method for response evaluation is reviewing actual response activities and interactions. Response reports are often available in the form of natural language text data. Analyzing a large number of such reports requires automated or semi automated solutions. To improve the trustworthiness of methods for this purpose, we empirically validate the reliability of three relation extraction methods that we used to construct interorganizational collaboration networks by comparing them against human-annotated ground truth (crisis-specific situational reports and tweets). For entity extraction, we find that using a combination of two off-the-shelf methods (FlairNLP and SpaCy) is optimal for situational reports data and one method (SpaCy) for tweets data. For relation extraction, we find that a heuristics-based model that we built by leveraging word co-occurrence and deep and shallow syntax as features and training it on domain-specific text data outperforms two state-of-the-art relation extraction models (Stanford OpenIE and OneIE) that were pre-trained on general domain data. We also find that situational reports, on average, contain less entities and relations than tweets, but the extracted networks are more closely related to collaboration activities mentioned in the ground truth. As it is widely known that general domain tools might need adjustment to perform accurately in specific domains, we did not expect the tested off-the-shelf tools to perform highly accurately. Our point is to rather identify what accuracy one could reasonably expect when leveraging available resources as-is for domain specific work (in this case, crisis informatics), what errors (in terms of false positives and false negatives) to expect, and how to account for that.

KEYWORDS

Collaboration networks, natural language processing, interorganizational collaboration, situational awareness

INTRODUCTION

Collaboration between organizations in large-scale crisis response events is critical as it ensures that response entities have the capacities and resources they need to meet the needs of vulnerable populations (Waugh Jr and Streib 2006; Ganji et al. 2019; Kapucu and Hu 2016). Specifically, collaboration structures during response reveal the extent to which shared situational awareness (SA) is achieved (Nofi 2000; Benson et al. 2010) and consequently how organizations collectively make decisions about response actions (Lee et al. 2012; Sarol, Dinh, and Diesner 2021; Sarol, Dinh, Rezapour, et al. 2020). Prior literature in emergency management has traditionally constructed organizational collaboration networks using surveys and interviews with emergency response personnel (Kapucu 2005; Ganji et al. 2019; Saoutal et al. 2015). While these methods can yield accurate information about actual collaborations that took place, the time and other resources needed for data collection limit the scalability of these methods (Guo and Kapucu 2015; Diesner and K. M. Carley 2005). As a result, a number of studies have leveraged unstructured text data such as police reports (Diesner and K. Carley 2010; Noori et al. 2016), news articles (Diesner, K. M. Carley, and Tambayong 2012; Sarol, Dinh, and Diesner 2021; Diesner and K. M. Carley 2004), and social media data (tweets) (Sheikh et al. 2017; Sarol, Dinh, Rezapour, et al. 2020; Sutton 2010) to extract information about organizational collaboration structures. The findings from these studies show that extracting crisis-related entities and their interactions from unstructured text data remains difficult due to a lack of crisis-specific a) training data and benchmarks (Sheikh et al. 2017; Nouali-TAboudjnent and Nouali 2011; Li et al. 2021) and b) knowledge about how to select methods and parameterization (Diesner 2012; Sarol, Dinh, and Diesner 2021; Diesner, Aleyasen, et al. 2014; Diesner 2013).

With the need for understanding collaboration structures during crisis response based on unstructured text data, this paper applies entity extraction and relation extraction to unstructured, natural language text data about hurricane events to improve our understanding of how to reliably obtain relevant network data that can than be used for network analysis, visualizations, simulations, etc (K. M. Carley, Diesner, et al. 2007). To this end, we compare the networks resulting from applying three different relation extraction techniques (two state-of-the-art methods and one newly developed heuristics-based method) to the same data sets (situational reports and tweets), and assessing the accuracy of each method by comparing the resulting networks against manually constructed ground truth network data about crisis-relevant collaboration. We recognize that while highly accurately performing off-the-shelf methods and models for entity and relation extraction exist (Honnibal and Montani 2017; Akbik et al. 2019; Angeli et al. 2015), they are not necessarily trained on crisis-specific text data. In response, we examine if and how methods and/ or models trained on crisis data (our heuristics-based model) differ from methods trained on domain-agnostic/ general domain text data. This leads to our two primary research questions:

RQ1: In what aspects is each of the three tested methods (heuristics-based, OneIE, OpenIE) accurate in detecting collaboration-related entities and relations, what errors do they make, how do the methods compare to each other?

RQ2: What do our findings imply for the practical use of these methods in the given domain?

Answers to these questions will help to design solutions that bring crises related network data extracted from texts closer to ground truth data, and accounting for remaining differences by at least being able to identify them. More specifically, this study contributes to the existing literature in crisis information systems in two ways: First, we identify and validate entities (of the type organizations, persons, geopolitical entities, national/religious/political groups, and locations) as well as relations between these entities that indicate collaboration activities that took place. Second, we further expand the knowledge about how choices for relation extraction methods and parameter settings influence networks we use for analysis, and consequently, impact our understanding of collaboration patterns that took place during response, which might be used as input to data-driven decision making and policy development or review.

RELATED WORK

Prior literature asserts that effective collaboration between organizations across governmental jurisdictions is a crucial success factor for large-scale crisis response (Comfort and Haase 2006; Kapucu 2005; Van Borkulo et al. 2005). A primary factor for effective collaboration between crisis response organizations is their shared SA about the main objectives as well as resources and actions needed to complete these objectives (Benson et al. 2010; Kurapati et al. 2013; Nofi 2000). A lack of SA may be due to several reasons, including inadequate technical infrastructure (Ganji et al. 2019), different organizational cultures and standards for response (Benssam et al. 2013), information overload (Hiltz and Plotnick 2013), and, most importantly, communication failures between organizational members in terms of not knowing what information is needed, from whom to request this information, and what agents or organization(s) have what resources (Saoutal et al. 2015; Ley et al. 2014; Bharosa et al. 2010; Weil, K. M. Carley, et al. 2006; Weil, Foster, et al. 2008).

We propose that taking a network perspective can help to solve this problem by making the observed response structures salient, visible, and analyzable, thereby allowing us to identify the entities that did collaborate during a response. Prior work has already leveraged organizational network analysis to examine patterns of collaboration among and between organizations at various levels of the government Kapucu (2009) and Kapucu and Garayev (2016). More specifically, network analysis has been used to analyze expected structures of collaboration based on federal guidelines (e.g. the Incident Command System) (Kapucu 2009) and to identify the key players (by calculating network centrality measures) to better understand who holds special roles in terms of power and influence in local-level response networks (Kapucu and Garayev 2016). While network analysis has been used to identify crisis-related content from unstructured text data (Sutton 2010; Diesner, K. M. Carley, and Tambayong 2012; Guo, Kapucu, and Huang 2021), network construction based on unstructured text data often requires labor-intensive methods such as sociometric surveys (Kapucu and Hu 2016) or content analysis to identify organizations and relations between them (Sutton 2010; Guo, Kapucu, and Huang 2021). These methods do not scale to analyzing large corpora. Other methods rely on unsupervised clustering to identify subgroups of organizations (Noori et al. 2016). These methods can return results that are not domain-relevant (Alam et al. 2020) if no training with crisis-related text data took place. Automated and semi-automated text mining methods exist, but entail limitations in terms of accuracy and/ or degree of automation (Diesner, K. M. Carley, and Tambayong 2012; Diesner 2015; K. M. Carley, Columbus, et al. 2008).

To address these limitations, we comparatively test three scalable relation extraction methods for their ability to *reliably* extract relevant relations, which we for the purpose of this project define as *collaboration*, between relevant entity types, e.g. organizations. Traditionally, text-based network construction has relied on word co-occurrences (user-defined window of words between a pair of entities) to extract relations (Danowski 1993; K. M. Carley 1997; K. M. Carley, Columbus, et al. 2008). However, this approach is susceptible to high false positive rates (i.e. retrieving incorrect relations (Diesner and K. M. Carley 2010; Diesner 2015)). To remedy this shortcoming, a number of relation extraction methods are available that leverage syntactic features such as shallow (Jiang and Diesner 2019) and deep parsing (Fundel et al. 2007; Xu et al. 2015) to classify and predict relations in semi- and un-structured texts. In addition to that, solutions such as Stanford's Open Information Extraction (Open IE) also leverage syntactic information (i.e. part-of-speech patterns and dependency trees) to extract triples, where a subject and an object are connected via a predicate (verb/verb phrase) (Angeli et al. 2015; Stanovsky et al. 2018; Gerner et al. 2002; Ogden and Richards 1925). Recent relation extraction methods have leveraged neural-network language models (Devlin et al. 2018; Dai and Le 2015) to learn more features (lexical, syntactic, and semantic) and improve overall accuracy (Wu and He 2019; Dai and Le 2015). In this context, OneIE (Lin et al. 2020) is a state-of-the-art method that was built by using a neural network-based model for supervised learning and extraction of both entities and relations at the same time. To date, OneIE outperforms existing solutions such as DyGIE++ (Wadden et al. 2019) and BERT-based event extraction models (Du and Cardie 2020).

In this paper, we apply state-of-the-art solutions to text data (situational reports and tweets) from the domain of crisis response to assess these solution's accuracy by comparing automatically extracted or predicted entities and relations to networks built from ground truth data, i.e., manual annotations of entities and relations. Specifically, we assess three models: (1) a heuristics-based model that we built by leveraging co-occurrence, shallow, and deep syntax as features, (2) Stanford's OpenIE (Qi et al. 2020; Angeli et al. 2015) prediction model that leverages part-of-speech tagging and dependency trees to construct relation triples, and (3) OneIE's joint entity and relation extraction model (Lin et al. 2020). In summary, constructing networks of entities that were mentioned in text data (Van Atteveldt 2008) to have collaborated in a disaster is relevant for reviewing disaster response and related policies ex post factum. However, this approach is also a highly specialized use case of entity and relation extraction, and most likely not the one that standard entity and relation extraction tools, such as CoreNLP and OneIE, were designed or fine tuned for. Therefore, we expect the congruence between the ground truth networks and the networks extracted by using general relation extraction tools to be lower than the tool's benchmark performances. For the same reason, we also expect the vast majority of mismatches not to be blamed on tools, but on the assumptions of the researchers conducting this work. In short, we want to find out to in what ways non-NLP-expert practitioners from an application domain can expect off-the-shelf tools to provide reliable information to them, and what errors they should expect and account for.

METHODOLOGY

Ground truth data

This study focuses on networks that encode or represent interorganizational collaboration during hurricane response as reported in text data. Specifically, we analyze four major hurricane events that made landfall within the continental United States, namely 2016 Hurricane Matthew, 2017 Hurricane Harvey, 2017 Hurricane Irma, and 2018 Hurricane Michael. We selected these events because they were federally-declared disasters based on the Stafford Act (Sobel and Leeson 2006) and thus warranted collaboration between organizations at the federal, state, local, and non-governmental level. We collected a total of 109 situational reports (SITREPs) and 28,050 Twitter posts (tweets) that were released two weeks before to two weeks after landfall of these four hurricanes. The ground truth data was constructed by (1) drawing a random sample of four SITREPs (approximately 8,000 words) and 1000 tweets (details in Figure 1), and (2) annotating entities (if any) and relations (if any) between entity pairs in our sample. Each of the three human coders marked up every sentence in the sample for entities. A sentence can contain zero, one, or many entities, and any two entities in a sentence may share a relationship or not. More specifically, based on the annotation schema we developed, an entity is defined as any named entity that is either a person (tag: PER), organization (tag: ORG), geopolitical entity (tag: GPE), national/religious/political groups (tag: NORP), or location (tag: LOC). The tag of each entity is then used as the entity type. We define a relationship as two entities being connected by a word or phrase (typically or noun or a verb) that indicates (to the trained human coders) an instance of *collaboration* between two entities, whereby two entities collaborate on a certain crisis-related task. To give an example, a sentence with two entities would be marked up for entity 1, entity 1 type, entity 2, entity 2 type, and a binary indication of a relation between entity 1 and entity 2 (relation exists or not).

For the SITREPs-based ground truth dataset, two coders engaged in a three-stage coding process (initial coding, reconciliation, and re-coding). They ultimately obtained an interrater agreement of 89.8%. For the tweets-based ground truth dataset, where we applied the same three-stage coding, three coders ultimately achieved an interrater agreement of 86.95%. Interrater agreement was assessed based on exact matches of (1) entity pairs (exact boundary matches, no partial matches counted, which is a strict condition), (2) entity types, and (3) existence of a relation (yes or no). The ground truth data was used for two purposes: First, to evaluate the two considered state-of-the-art relation extraction tools against this gold standard. Second, to computationally identify heuristics for locating relevant entities and relations, and discriminating those from irrelevant words, entities, and relations, and then using these heuristics as features for training a model for predicting or extracting entities and relations (pipeline discussed in next section).

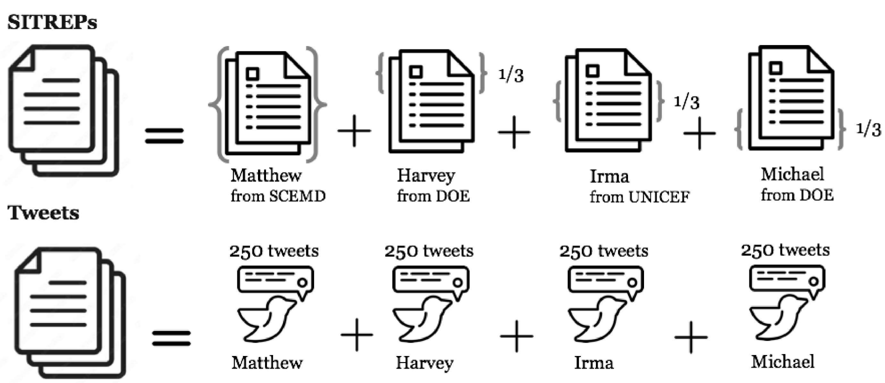

Figure 1. Ground Truth Sample

Relation extraction methods

We assess the accuracy of the three outlined relation extraction methods in terms of locating and classifying entities (n-ary classification) and relations (binary classification) that represent nodes and edges in collaboration networks during hurricane responses. In particular, we develop a heuristics-based method for entity and relation extraction and compare the performance of this method to two existing state-of-the-art relation extraction methods (OneIE (Lin et al. 2020) and Open Information Extraction (OpenIE, (Stanovsky et al. 2018)). The sections below discuss each method in detail.

Heuristics-based model

The model entails (1) named entity detection and labeling based on a combination of two off-the-shelf models (SpaCy and FlairNLP), followed by (2) relation extraction via a combination of optimal features based on entity co-occurrence as well as shallow and deep syntax parsing. By optimal we here mean resulting in the highest accuracy in our annotated ground truth data. The named entity labeling process entails finding optimal settings for applying SpaCy and FlairNLP. As shown in Figure 4, we found that the optimal setting is to choose entities that are labeled by both Flair and SpaCy (i.e. Flair + SpaCy) for the SITREPs-based ground truth, and entities labeled solely by SpaCy for the tweets-based ground truth. We speculate this result is due to differences in Flair and SpaCy's pre-trained models (even though both were trained on OntoNotes 5.0 (Weischedel et al. 2012)) that may deem one method more suitable for tagging tweets data than another. For the relation extraction task, we identified the following optimal features: (1) top three co-occurrence-based window sizes, (1) top three part-of-speech patterns from constituency trees, (3) top three dependency-based window sizes, and (4) top three dependency patterns between every two entities. For co-occurrence, window size indicates a number of words in between two entities, not including the entities themselves. For example, a co-occurrence window size of zero represents adjacent entities, with zero words in between them. For dependency, window size represents the shortest path length between two entities in the dependency parse tree. For instance, a dependency window size of zero represents two entities connected by a direct dependency path from one entity to another. The top three instances of each feature are shown in Tables 1–3. With these features, we trained a Support Vector Machine (SVM) model with a 80%-20% train-and-test split, along with 10-fold cross-validation to evaluate the performance of the heuristics-based model. We acknowledge that building a model with these "optimal" features can lead to overfitting when applied to new and unseen data, even after k-fold cross validation was applied, such that further calibration with other data from the same domain might be beneficial. The point here is mainly to compare a domain-specific solution to general domain alternatives.

Table 1. Top co-occurrence-based (left) and dependency-based (right) window sizes of relations present in SITREPs-based and tweets-based ground truth data

	# of relations with **co-occurrence-based** window size in:		# of relations with **dependency-based** window size in:	
Window size	SITREPs	Tweets	SITREPs	Tweets
0	10 (1.414%)	1 (0.124%)	62 (8.769%)	4 (0.498%)
1	**134 (18.953%)**	**127 (15.796%)**	**170 (20.045%)**	67 (8.333%)
2	**102 (15.134%)**	93 (11.567%)	70 (9.9%)	99 (12.313%)
3	**97 (13.72%)**	**126 (15.7%)**	**86 (12.164%)**	145 (18%)
4	85 (12.022%)	**113 (14.054%)**	**72 (10.183%)**	158 (19.652%)
5	71 (10.042%)	102 (12.687%)	62 (8.769%)	104 (12.935%)
6	57 (8.062%)	78 (9.701%)	47 (6.648%)	69 (8.582%)
7	55 (7.779%)	68 (8.458%)	46 (6.506%)	34 (4.229%)
8	48 (6.789%)	55 (6.841%)	32 (4.526%)	26 (3.234%)
9	43 6.082%)	41 (5.099%)	23 (3.253%)	3 (0.373%)

Table 2. Top three part-of-speech patterns between two entities in SITREPs-based and tweets-based ground truth

Constituency pattern	# of relations in **SITREPs** ground truth	Constituency pattern	# of relations in **tweets** ground truth
1-8 PROPNs	29.8%	1-3 PROPNs	34.415%
CCONJ	3.9%	VERB-ADP	12.553%
ADP	1.6%	NOUN-ADP	6.488%

Table 3. Top three dependency patterns between two entities in SITREPs-based and tweets-based ground truth

Dependency pattern	# of relations in **SITREPs** ground truth	Dependency pattern	# of relations in **tweets** ground truth
2-16 PROPNs	48.949%	VERB-ADP	22.003%
VERB-ADP-PROPN	3.704%	2-4 PROPNs	14.81%
VERB-NOUN	1.401%	NOUN-ADP	6.347%

OneIE

OneIE (Lin et al. 2020) extracts relation(s) from a sentence by encoding each word, identifying and classifying entity mentions and event triggers, and decoding to find the graph with the highest global score. The context of each input sentence (and the words within a sentence) is determined using a pre-trained BERT encoder (Devlin et al. 2018). Entity mentions and event triggers are identified using a feed-forward neural network, meaning that the connection(s) between entities do not form a loop. The outputs from that become the nodes in our resulting network. Label scores are calculated by averaging each word representation and employing another feed-forward network. Finally, a beam search-based decoder (sample multiple sequences at once then predict iteratively) was implemented to find the optimal network with the overall highest global score.

OneIE is trained based on the Automatic Content Extraction (ACE 2005) dataset. ACE 2005 contains different data genres, including newswire data, broadcast news and conversation, weblogs, discussion forums, and telephone speeches (Walker et al. 2006). As this study focuses on English text data, we leverage the ACE2005 English training dataset, which includes annotations of entity mentions, entity mention types (named entities only), relations, and events. Each entity mention is tagged with one of seven entity types (PER, ORG, GPE, LOC, vehicles (VEH), facilities (FAC), and weapons). Each relation between an entity pair can be one of six types, namely person-social (PER entity pair), physical (PER/FAC/LOC/GPE entity pair), general affiliation (PER-PER/LOC/GPE/ORG entity pair), part-whole (FAC/LOC/GPE/ORG entity pair), org-affiliation (PER/ORG/GPE entity pair), and artifact (PER/ORG/GPE-FAC entity pair). We consider all relation types in this study, and created an additional category that groups all tagged relations as "relation" edges and those without relation tags as "no relation". The training data also contains relations coded into event types (n=33) and argument roles (n=22), which we did not consider as they are beyond the scope of this study.

OpenIE

OpenIE (Stanovsky et al. 2018) is different from OneIE in that relations extracted are not limited to predefined entity types and relationship types based on training data (Etzioni et al. 2008). Instead, OpenIE leverages part-of-speech (POS) tagging and syntactic parsing to extract triples (subject, predict, object) from a sentence. OpenIE is suitable for the scope of this study as the output includes entity-relation-entity information, and is thus suitable for network construction. We leverage the Stanford CoreNLP's implementation of OpenIE (Qi et al. 2020) to extract triples from our SITREPs and tweets data, where each entity is tagged based on CoreNLP's fine-grained NER annotator, which includes 24 entity types (e.g. PER, ORG, LOC, and fine-grained categories such as CITY, RELIGION, CAUSE_OF_DEATH). The relation for each subject-predicate-object triple has no predefined label, but is rather a verb/verb phrase that lies in between two entities. An example of relation triples extracted by OpenIE is shown in Figure 2, where OpenIE outputs five triples from one input sentence.

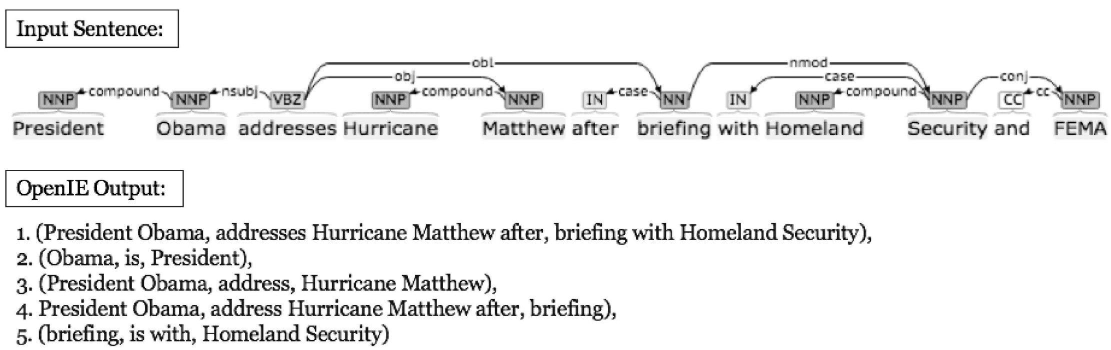

Figure 2. OpenIE output for an input sentence in Hurricane Matthew's SITREPs dataset

EVALUATION AND NETWORK ANALYSIS RESULTS

In this section, we compare the resulting networks from each entity and relation extraction methods to the ground truth. We start with evaluating the performance of each method in terms of precision, recall, and the F1 measure, where F1 was calculated based on whether an entity pair was found to have a relation that is also present in the ground truth data or not. While each of the two entities in a link must match the ground truth in boundary and type, the order of occurrence of these entities in a sentence does not matter (e.g. UN <-> EPA is the same as EPA <-> UN). Any partial entity matches are not considered a match. For the evaluation of entity extraction, we require a perfect match in entity name and entity type with the ground truth. For the evaluation of relation extraction, we require a perfect match in the linked entities (as in prior sentence) and in the binary classification of the existence of a relationship between an entity pair.

The performance results for the three entity extraction methods are shown in Table 4, and the results for the three relation extraction methods in Table 5. Among the three entity and relation extraction methods, our heuristics-based entity and relation extraction methods yield the highest performance (in terms of F1 measure). Among the other two methods, OneIE consistently outperforms OpenIE for both entity and relation extraction. Across the two ground truth datasets, all methods yield better performance on SITREPs than tweets. In particular, OpenIE-based entity and relation extraction on tweets yields an F1 measure of 0, primarily because there is no match in entity type (e.g. Haiti:*COUNTRY* (OpenIE) versus Haiti:*GPE* (ground truth). Dropping the entity type match requirement, we find that the F1 measure for OpenIE would increase to 0.912 (precision: 0.899, recall: 0.924), and relation extraction accuracy would also improve (though minimally: F1: 0.041, precision: 0.025, recall: 0.118).

Table 4. Performance of Entity Extraction methods

Model	SITREPs			Tweets		
	Precision	Recall	F1	Precision	Recall	F1
OneIE	0.163	0.269	0.203	0.100	0.156	0.122
OpenIE	0.0006	0.0006	0.0006	0	0	0
Heuristic Method (newly developed)						
Flair	0.840	0.718	0.774	0.838	0.337	0.481
SpaCy	0.772	**0.799**	0.678	0.730	**0.576**	**0.644**
Flair + SpaCy	**0.909**	0.722	**0.805**	**0.895**	0.394	0.547
Flair or SpaCy	0.777	0.570	0.658	0.751	0.444	0.558

Table 5. Performance of Relation Extraction methods

Model	SITREPs			Tweets		
	Precision	Recall	F1	Precision	Recall	F1
OneIE	0.408	0.498	0.449	0.051	0.110	0.070
OPENIE	0.033	0.135	0.053	0	0	0
Heuristic Method (newly developed)	**0.928**	**0.913**	**0.920**	**0.928**	**0.820**	**0.870**

Next, we compared the networks resulting from each method's entity and relation extraction output (shown in Figures 4-6) to the ground truth networks (shown in Figure 3) in terms of (1) network structure and (2) amount of overlaps in entities and relations. A visual examination indicates that the networks generated by the heuristics-based method resemble the ground truth networks most closely based on the overall structure of entities and relations between them, and network density. Specifically for the ground truth and heuristics-based SITREPs networks, there is a dense cluster of entities (names of counties) surrounded by smaller subgroups and triads. This shows that the heuristics-based SITREPs network accurately captures the active collaboration between county-level emergency operation centers that are recorded in the ground truth. In contrast to that, the tweets-based networks show no prominent clusters, but small subset of nodes that are central in both networks (e.g. hurricane names and state names). This finding shows that the heuristics-based tweets network fails to capture the collaboration patterns present in the ground truth network. OneIE's and Stanford OpenIE's based networks are notably different from the ground truth networks: Among the SITREPs-based networks, the one generated by OneIE is least dense, thus does not capture the dense cluster of connections between local-level entities in the ground truth data. Similarly, OneIE's tweets-based network is the least dense tweets network, however, the central nodes detected (i.e. state names such as Florida and Texas) are the same as those in the ground truth network. The networks extracted by OpenIE are distinctive from the ground truth and other empirical networks in that they are notably larger and contain more clusters.

Table 6. Network descriptives for ground truth networks (SITREPs-based and tweets-based) and networks resulting from heuristics-based, OneIE-based, and OpenIE-based entity and relation extraction methods. Preprocessing includes removing self-loops and isolates

	GT		Heuristics		OneIE		OpenIE	
Network measures	SITREPs	Tweets	SITREPs	Tweets	SITREPs	Tweets	SITREPs	Tweets
# of Nodes	300	531	183	892	166	377	1,431	2,600
# of Edges	1,304	1,306	1,019	7,001	210	418	1,737	3,371
Deg. Centralization	8.693	4.919	11.198	15.697	2.530	2.217	2.428	2.593
Density	0.029	0.009	0.062	0.018	0.015	0.006	0.002	0.001
Global Clustering	0.454	0.153	0.509	0.749	0.042	0.046	0.004	0.007
Average Path Length	3.617	3.485	3.236	2.892	5.737	4.559	8.964	5.283
Size of Largest Component	1,218	1,267	990	6,915	123	287	938	2,360

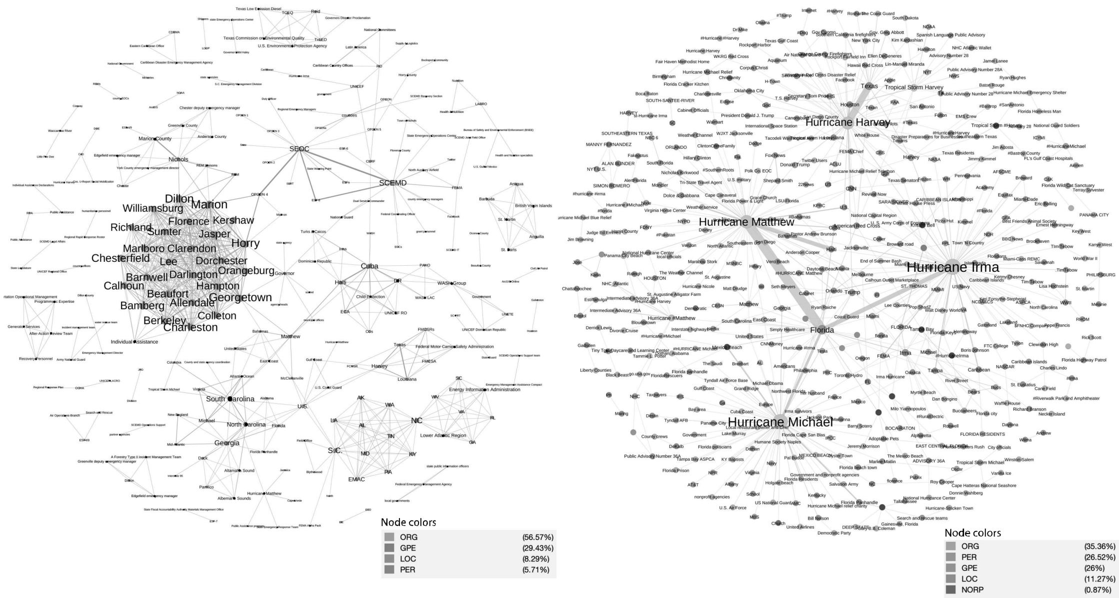

Figure 3. SITREPs-based (left) and tweets-based (right) networks generated with ground-truth data

Common network analytic metrics are shown in Table 6. Looking at the three relation extraction methods, the size of the heuristics-based networks is most similar to the ground truth, followed by OneIE-based networks, though those are slightly smaller with notably fewer edges. OpenIE-based networks are the largest, followed by heuristics-based networks and OneIE-based networks. The tweets-based networks are, on average, larger (average size=1,100 edges) than the SITREPs-based networks (average size=520 edges), which is different from the ground truth networks were SITREPs and tweets networks are similar in size (1,304 and 1,306, respectively, although tweets-based network has more nodes). Similar to the ground truth networks, networks from all three methods have low densities, indicating that the actual number of relations between entities is much lower than expected. In terms of topological characteristics, the heuristics-based and ground truth-SITREPs networks are the only networks with high global clustering and short average path lengths. This combination of features is an indicator for a small-world network (Watts and Strogatz 1998), where on average, an entity has many immediate connections and a few distant connections. The remaining networks show relatively low global clustering and high average path length, suggesting that these networks are not behaving like small-world networks.

Beyond the structural network differences, tweets-based networks resulting from the three methods contain more diversified entities than SITREPs-based networks, such as celebrity names (e.g. Vanilla Ice, Donnie Wahlberg), political parties (e.g. Republicans, Democratic Party), and mentions of sports teams (e.g. [Chicago] Bears, [Tampa Bay] Buccaneers), some of which may not be related to crisis collaboration activities. Though smaller in network size, SITREPs-based networks extracted from the three methods contain more crisis-related organizations at multiple levels of the government (e.g. Red Cross, Bay County Emergency Operations Center) and individuals (e.g. Gov Cuomo, SERT personnel) involved in response collaborations. Both SITREPs-based and tweets-based networks overall contain specific location names (e.g. Tyndall Air Force Base, Blythewood Field Office), which are helpful to identify areas of collaboration activities, and tweets-based networks contain more generic locations (e.g. Airports, Atmosphere, Cities) than SITREPs-based networks.

In this section, we report on the types of entities and relations contained in the networks extracted with the three tested methods. OpenIE finds more entity types that differ from the ground truth networks than the other two methods, despite the fact that OpenIE's NER model has fewer entity types than heuristics-based's NER model (12-class versus 18-class, respectively). The most prominent entity types detected by OpenIE in the SITREPs-based network, i.e. *numbers* (26.21%) and *date* (9.22%), may not provide as much information about collaboration activities as *organizations* (20.13%) and location (6.92%). For the tweets-based networks, OpenIE detects a substantial number of *URLs* (56.24%), followed by *organizations* (9.31%) and *persons* (7.51%). As for top relationship types detected, OpenIE's SITREPs-based network in dominated by "were in" (7.58%), "is providing" (6.44%), and 'was on" (4.29%). Only "is providing" is potentially indicative of some collaboration activity between two entities, while the other two types may be indicative of physical co-location of agents and/ or resources. In OpenIE's tweets-based network, the top relationship types are "is" (5.52%), "is in" (3.6%), and ''has" (1.99%). The top relations in

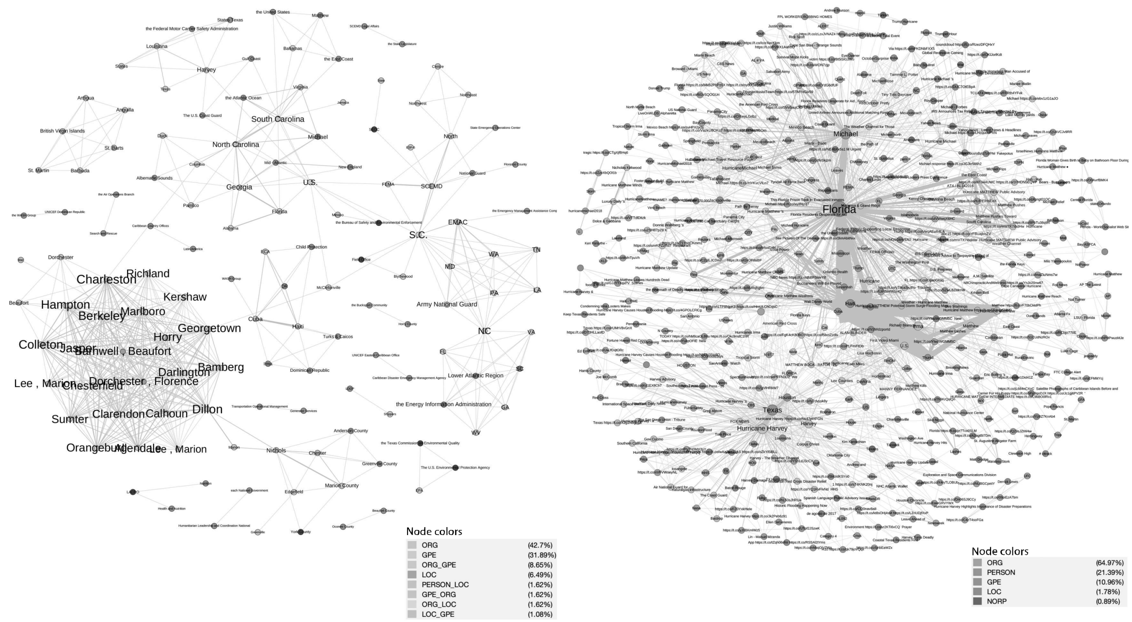

Figure 4. SITREPs-based (left) and tweets-based (right) networks generated with <u>heuristics-based</u> relation extraction method

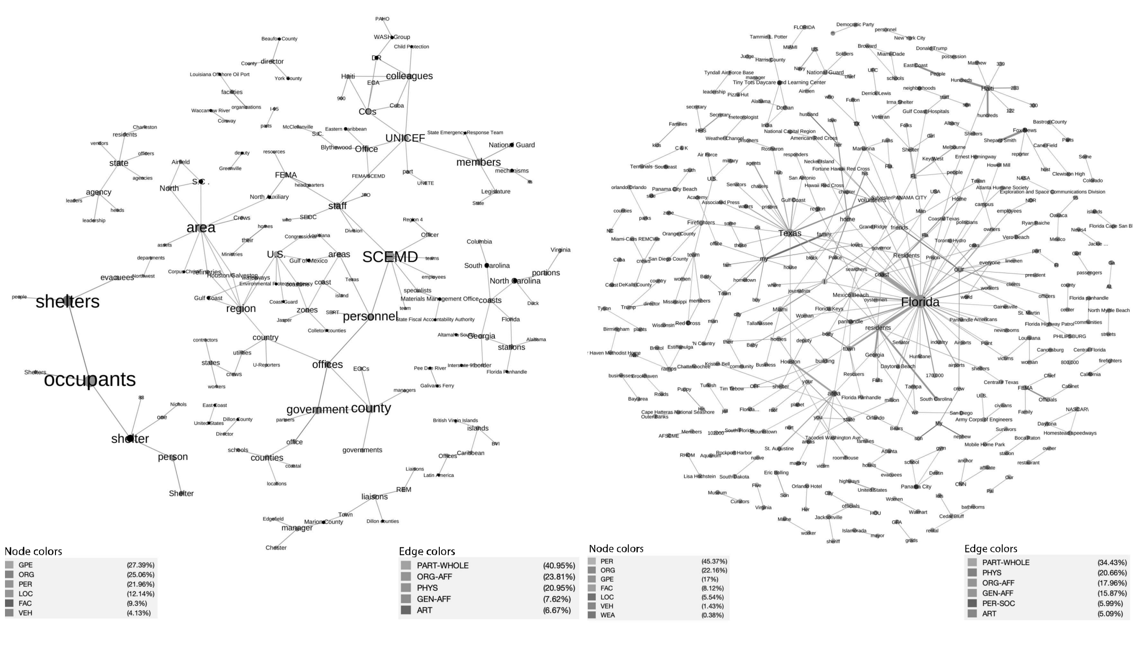

Figure 5. SITREPs-based (left) and tweets-based (right) networks generated with <u>OneIE</u> relation extraction method

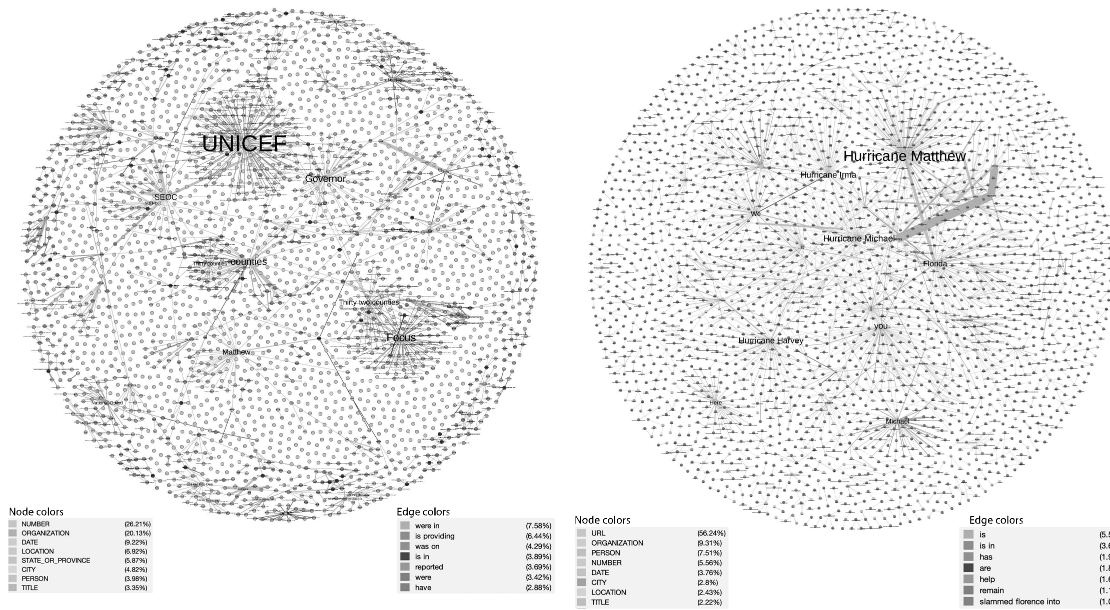

Figure 6. SITREPs-based (left) and tweets-based (right) networks generated with Stanford OpenIE relation extraction method

OpenIE's tweets-based network, however, is not representative of the entire (long-tail) distribution of relations detected. Hence, a larger number of relations detected in tweets compared to SITREPs show that the language used in tweets is more diversified than that in SITREPs.

One-IE's tweets-based network contain more types of entity and relations than the SITREPs-based network, with *persons* (45.37%), *organizations* (22.16%), and *geopolitical entities* (17%) being prominent entities. However, these entities are not exact matches with those mentioned in the ground truth tweet network, resulting in low performance (F1 for NER: 0.112). Between these entities, most relations are of the types "part-whole" (34.43%), "physical" (20.66%), and "organization-affiliation" (17.96%). These relations are also not explicitly present in the ground truth networks because the ground truth only contains a binary relation/no relation. OneIE's SITREPs-based network also has the same top three entity types (*geopolitical entities* (27.39%), *organizations* (25.06%), and *persons* (21.96%)) as the tweets-based network. The entity pairs are also frequently connected via "part-whole" (40.95%), "organization-affiliation" (23.81%), and "physical" (20.95%) relationship types.

The heuristics-based networks contain the least amount of entity types compared to the other two methods (for SITREPs, four entity types detected as opposed to six types detected in OneIE and eight types detected in OpenIE; for tweets, five entity types detected as opposed to seven types detected in OneIE and eight types detected in OpenIE), however, the relations predicted have the highest performance compared to the other methods. In the heuristics-based SITREPs network, a majority of entities are tagged as *organizations* (56.57%), followed by *geopolitical entities* (29.43%), and *locations* (8.29%). There are no entities with the *national, religious, or political groups* tag detected in this network. For the tweets-based network, entities detected are also primarily *organizations* (35.36%), *persons* (26.52%), and *geopolitical entities* (26%). *National, religious, or political groups* entities are present, but minimal (0.87%, e.g. "Democratic", "Texans").

DISCUSSION AND FUTURE WORK

Our first research question asked what collaboration-related entities and relations each of the three tested methods (heuristics-based, OneIE, OpenIE) gets right (in comparison to the ground truth), what errors each method makes, and how the tested methods compare to each other. The variability in performance results and networks generated by these methods further confirms that methodological choices about node and edge detection impact and potentially distort how we perceive the evolution of collaboration among various actors during crisis response. Specifically, the entity and relation extraction results confirm that interorganizational collaboration is closest to ground truth when using our heuristic-based model. However, our solution might overfit to domain data and needs further calibration. Furthermore, models trained with supervised learning (i.e. heuristics-based and OneIE methods) and unsupervised approaches (i.e. OpenIE) yield different results, indicating that training and testing are crucial to ensure that relation extraction is applicable to crisis-related content. Furthermore, OpenIE's low performance in terms of both entity type labeling and, consequently, relation extraction may suggest behavior early on in the syntactic parsing and part-of-speech tagging pipeline that drives accuracy away from ground truth resemblance, which in turn influences the accuracy of the extraction of triples (Gashteovski 2020). Another possible explanation might be that OpenIE's definition and operationalization of a relation differs from what we did in the ground truth, and thus logically leads to low overlaps in entities and relations extracted. OneIE, while it performs better than OpenIE for this particular application scenario, still contains high false positive and false negative rates, specifically higher number of false positives than false negatives (precision lower than recall). Given that OneIE was pre-trained with non-crisis related benchmark dataset (i.e. ACE2005 dataset[1], the results suggest that training with domain-specific data may help boost performance. The networks generated with OneIE, however, do contain entities and relations that are relevant to crisis response (shown in Figure 5), and thus a joint entity-relation extraction pipeline (where NER and RE are performed simultaneously to minimize errors cascading from one task to another) is useful for this context. It is furthermore imaginable that the employed tools do find false positives, i.e., entities and/ or relations relevant to crisis response but not contained in the ground truth, which would require further investigation to test for that.

Our second research question is about the implications of our findings for bringing the tested methods into practical application scenarios in crisis research. We have shown that the application of state-of-the-art relation extraction methods requires domain-adaptation with crisis-specific training data. Furthermore, the choices for what entity and relation extraction method or combination of methods to use, how to configure them, and what types of entity and relations to consider in the scope of a crisis response network do influence the networks generated. Last but not least, it is important to determine whether a research study's definition of a relation is aligned with the method's operationalization of a relation. For instance, our heuristics-based method focuses on the binary detection of relation, while OneIE and OpenIE extract multiple relation types which may be helpful to further explore and categorize the content of relations. We also demonstrated that a heuristics-based model trained with crisis-specific text data can leverage text features (including word co-occurrence and constituency and dependency parse trees) that more accurately reflect crisis-specific language use than general-domain models (i.e. OneIE and OpenIE). Understanding the language used in crisis text data is thus necessary to distill relevant information about the entities involved and the interactions that took place between them during response.

Our work is driven by the assumption that effective collaboration relies on adequate shared SA of the conditions and needs surrounding a crisis (U.S. Department of Homeland Security 2008; Nofi 2000; Benson et al. 2010) to inform decision-making and guide response actions. However, prior literature also finds that information exchange between organizations, which is necessary to achieve shared SA, remains difficult (Bjurling and Hansen 2010; Saoutal et al. 2015). One reason for this difficulty is the lack of formal information pipelines for connecting requesting and responding organizations to various crisis-related resources and services (Lee et al. 2012; Sarol, Dinh, Rezapour, et al. 2020). Consequently, the lack of collaboration between requesting and responding organizations may negatively influence the overall ability of members within an interorganizational network to dynamically interact with each other (Benssam et al. 2013; Kurapati et al. 2013) and to address crisis-related needs (Ganji et al. 2019). To address this shortcoming, our analysis of the collaboration dynamics based on situational reports and tweets surrounding actual hurricane response shows that organizations and persons at various levels of the government (e.g. FEMA, Florida Highway Patrol, Bay County Operations Center) did work together with non-governmental entities (e.g. UNICEF, United Methodist Committee on Relief, American Society for the Prevention of Cruelty to Animals). While doing a substantive analysis of these dynamics is outside the scope of this paper and part of our future work, the work shown in this paper enables the collection or construction of data needed for such analyses. We also showed that domain-adaptation of entity and relation extraction models is important to reliably extract crisis-related collaboration information from text data. In future work, we will also continue to examine existing structures of crisis collaboration to identify areas of improvement for crisis response policy planning. In terms of methodological improvements, we plan to expand our ground truth corpus to include more texts from different crisis events to cover the language used in crisis reporting and communication more broadly and thereby advancing the generalizability of computational, domain-specific information extraction solutions.

ACKNOWLEDGEMENT

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-CIRC01. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. We appreciate helpful feedback and continuous support from the Critical Infrastructure Resilience Institute (CIRI) and U.S. Coast Guard.

[1] https://catalog.ldc.upenn.edu/LDC2006T06

REFERENCES

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). "FLAIR: An easy-to-use framework for state-of-the-art NLP". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59.

Alam, F., Ofli, F., and Imran, M. (2020). "Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria". In: *Behaviour & Information Technology* 39.3, pp. 288–318.

Angeli, G., Premkumar, M. J. J., and Manning, C. D. (2015). "Leveraging linguistic structure for open domain information extraction". 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)*, pp. 344–354.

Benson, A. L., Biggers, K., Wall, J., and Haselkorn, M. P. (2010). "Adaptive development of a common operating environment for crisis response and management." In: *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Benssam, A., Nouali-Taboudjemat, N., and Nouali, O. (2013). "Towards an It-based platform for disaster risks management in Algeria." In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Bharosa, N., Lee, J., and Janssen, M. (2010). "Challenges and obstacles in sharing and coordinating information during multi-agency disaster response: Propositions from field exercises". In: *Information Systems Frontiers* 12.1, pp. 49–65.

Bjurling, B. and Hansen, P. (2010). "Contracts for information sharing in collaborative networks." In: *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Carley, K. M. (1997). "Network text analysis: The network position of concepts". In: *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* 4, pp. 79–100.

Carley, K. M., Columbus, D., Diesner, J., and DeReno, M. (2008). *AutoMap user's guide. CMU-ISR-08-123*. Tech. rep. Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Carley, K. M., Diesner, J., Reminga, J., and Tsvetovat, M. (2007). "Toward an interoperable dynamic network analysis toolkit". In: *Decision Support Systems* 43.4, pp. 1324–1347.

Comfort, L. K. and Haase, T. W. (2006). "Communication, coherence, and collective action: The impact of Hurricane Katrina on communications infrastructure". In: *Public Works management & policy* 10.4, pp. 328–343.

Dai, A. M. and Le, Q. V. (2015). "Semi-supervised sequence learning". In: *Advances in neural information processing systems* 28.

Danowski, J. A. (1993). "Network analysis of message content". In: *Progress in communication sciences* 12, pp. 198–221.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Diesner, J., Aleyasen, A., Mishra, S., Schecter, A., and Contractor, N. (2014). "Comparison of Communication Networks built from explicit and implicit data". In: *Computational Approaches to Social Modeling (CHASM 2014), ACM WebScience Conference*.

Diesner, J. and Carley, K. (2010). "Relational methods for analyzing and preventing crime (German: Relationale Methoden in der Erforschung, Ermittlung und Prävention von Kriminalität)". In: *Handbook of Network Research*, pp. 725–738.

Diesner, J. (2012). *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts*. Tech. rep. Carnegie Mellon University Institute of Software Research.

Diesner, J. (2013). "From texts to networks: detecting and managing the impact of methodological choices for extracting network data from text data". In: *KI-Künstliche Intelligenz* 27.1, pp. 75–78.

Diesner, J. (2015). "Words and networks: How reliable are network data constructed from text data?" In: *Roles, Trust, and Reputation in Social Media Knowledge Markets*. Springer, pp. 81–89.

Diesner, J. and Carley, K. M. (2004). "Using network text analysis to detect the organizational structure of covert networks". In: *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference*. Vol. 3. NAACSOS Pittsburgh.

Diesner, J. and Carley, K. M. (2005). "Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis". In: *Causal mapping for research in information technology*. IGI Global, pp. 81–108.

Diesner, J. and Carley, K. M. (2010). "Extraktion relationaler Daten aus Texten". In: *Handbuch Netzwerkforschung*. Springer, pp. 507–521.

Diesner, J., Carley, K. M., and Tambayong, L. (2012). "Extracting socio-cultural networks of the Sudan from open-source, large-scale text data". In: *Computational and Mathematical Organization Theory* 18.3, pp. 328–339.

Du, X. and Cardie, C. (2020). "Event extraction by answering (almost) natural questions". In: *arXiv preprint arXiv:2004.13625*.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). "Open information extraction from the web". In: *Communications of the ACM* 51.12, pp. 68–74.

Fundel, K., Küffner, R., and Zimmer, R. (2007). "RelEx—Relation extraction using dependency parse trees". In: *Bioinformatics* 23.3, pp. 365–371.

Ganji, A., Wilson, T., Saveli, S., Dailey, D., and Haselkorn, M. P. (2019). "Cause and Effect: A Qualitative Analysis of Obstacles to Information Sharing During a Regional Disaster Exercise." In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Gashteovski, K. (2020). "Compact open information extraction: methods, corpora, analysis". In.

Gerner, D. J., Schrodt, P. A., Yilmaz, O., and Abu-Jabr, R. (2002). "Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions". In: *International Studies Association, New Orleans*.

Guo, X. and Kapucu, N. (2015). "Network performance assessment for collaborative disaster response". In: *Disaster Prevention and Management* 24.2, pp. 201–220.

Guo, X., Kapucu, N., and Huang, J. (2021). "Examining resilience of disaster response system in response to COVID-19". In: *International Journal of Disaster Risk Reduction* 59, p. 102239.

Hiltz, S. R. and Plotnick, L. (2013). "Dealing with information overload when using social media for emergency management: Emerging solutions." In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pp. 823–827.

Honnibal, M. and Montani, I. (2017). "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* 7.

Jiang, M. and Diesner, J. (2019). "A constituency parsing tree based method for relation extraction from abstracts of scholarly publications". In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pp. 186–191.

Kapucu, N. (2005). "Interorganizational coordination in dynamic context: Networks in emergency response management". In: *Connections* 26.2, pp. 33–48.

Kapucu, N. (2009). "Interorganizational coordination in complex environments of disasters: The evolution of intergovernmental disaster response systems". In: *Journal of Homeland Security and Emergency Management* 6.1.

Kapucu, N. and Garayev, V. (2016). "Structure and network performance: Horizontal and vertical networks in emergency management". In: *Administration & Society* 48.8, pp. 931–961.

Kapucu, N. and Hu, Q. (2016). "Understanding multiplexity of collaborative emergency management networks". In: *The American Review of Public Administration* 46.4, pp. 399–417.

Kurapati, S., Kolfschoten, G. L., Verbraeck, A., Corsi, T. M., and Brazier, F. (2013). "Exploring shared situational awareness in supply chain disruptions". In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM).

Lee, G., Oh, N., and Moon, I.-C. (2012). "Modeling and simulating network-centric operations of organizations for crisis management". In: *Proceedings of the 2012 Symposium on Emerging Applications of M&S in Industry and Academia Symposium*, pp. 1–8.

Ley, B., Ludwig, T., Pipek, V., Randall, D., Reuter, C., and Wiedenhoefer, T. (2014). "Information and expertise sharing in inter-organizational crisis management". In: *Computer Supported Cooperative Work (CSCW)* 23.4, pp. 347–387.

Li, Y., Liao, D., Li, J., and Ji, W. (2021). "Automated generation of interorganizational disaster response networks through information extraction". In: *arXiv preprint arXiv:2103.00287*.

Lin, Y., Ji, H., Huang, F., and Wu, L. (2020). "A joint neural model for information extraction with global features". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7999–8009.

Nofi, A. A. (2000). *Defining and measuring shared situational awareness*. Tech. rep. Center For Naval Analyses Alexandria VA.

Noori, N. S., Wolbers, J., Boersma, K., and Vilasıés-Cardona, X. (2016). "A Dynamic Perspective of Emerging Coordination Clusters in Crisis Response Networks." In: *Proceedings of the 13th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Nouali-TAboudjnent, N. and Nouali, O. (2011). "Knowledge extraction by Internet monitoring to enhance crisis management". In: *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Vol. 1.

Ogden, C. K. and Richards, I. A. (1925). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Vol. 29. Harcourt, Brace.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). "Stanza: A Python natural language processing toolkit for many human languages". In: *arXiv preprint arXiv:2003.07082*.

Saoutal, A., Matta, N., Cahier, J.-P., et al. (2015). "Approach to support Situational Awareness within Inter-Organizational Collaboration in Crisis Response." In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Citeseer.

Sarol, M. J., Dinh, L., and Diesner, J. (2021). "Variation in Situational Awareness Information due to Selection of Data Source, Summarization Method, and Method Implementation". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15, pp. 597–608.

Sarol, M. J., Dinh, L., Rezapour, R., Chin, C.-L., Yang, P., and Diesner, J. (2020). "An Empirical Methodology for Detecting and Prioritizing Needs during Crisis Events". In: *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4102–4107.

Sheikh, Z., Masood, H., Khan, S., and Imran, M. (2017). "User-Assisted Information Extraction from Twitter During Emergencies." In: *Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Sobel, R. S. and Leeson, P. T. (2006). "Government's response to Hurricane Katrina: A public choice analysis". In: *Public Choice* 127.1, pp. 55–73.

Stanovsky, G., Michael, J., Zettlemoyer, L., and Dagan, I. (2018). "Supervised open information extraction". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 885–895.

Sutton, J. N. (2010). "Twittering Tennessee: Distributed networks and collaboration following a technological disaster". In: *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Citeseer.

U.S. Department of Homeland Security (2008). *National Response Framework*.

Van Atteveldt, W. H. (2008). "Semantic network analysis: Techniques for extracting, representing, and querying media content". In.

Van Borkulo, E., Scholten, H. J., Zlatanova, S., and Brink, A. van den (2005). "Decision making in response and relief phases". In: *Geo-information for Disaster Management, First International Symposium on Geo-information for Disaster Management, Delft, 31-23 March 2005*, pp. 47–53.

Wadden, D., Wennberg, U., Luan, Y., and Hajishirzi, H. (2019). "Entity, relation, and event extraction with contextualized span representations". In: *arXiv preprint arXiv:1909.03546*.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). "ACE 2005 multilingual training corpus". In: *Linguistic Data Consortium, Philadelphia* 57, p. 45.

Watts, D. J. and Strogatz, S. H. (1998). "Collective dynamics of 'small-world'networks". In: *nature* 393.6684, pp. 440–442.

Waugh Jr, W. L. and Streib, G. (2006). "Collaboration and leadership for effective emergency management". In: *Public administration review* 66, pp. 131–140.

Weil, S. A., Carley, K. M., Diesner, J., Freeman, J., and Cooke, N. J. (2006). *Measuring situational awareness through analysis of communications: A preliminary exercise*. Tech. rep. APTIMA INC WOBURN MA.

Weil, S. A., Foster, P., Freeman, J., Carley, K., Diesner, J., Franz, T., and Gorman, J. (2008). "Converging approaches to automated communications-based assessment of team situation awareness". In: *Macrocognition in Teams. Theories and Methodologies*, pp. 277–300.

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., et al. (2012). "Ontonotes release 5.0". In: *LDC2012T19, Philadelphia, Penn.: Linguistic Data Consortium*.

Wu, S. and He, Y. (2019). "Enriching pre-trained language model with entity information for relation classification". In: *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 2361–2364.

Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., and Jin, Z. (2015). "Classifying relations via long short term memory networks along shortest dependency paths". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1785–1794.