Events and Stories: NLP toward Secure Software Engineering

2021 SoS Summer Quarterly

Hui Guo and Munindar P. Singh

July 13th, 2021

Info: https://hguo5.github.io/phddefense/

Email: hguo5@ncsu.edu, mpsingh@ncsu.edu

NC STATE UNIVERSITY

Introduction

Motivation

Natural language text in software engineering

Stories: prevalent in NL text

Rich information about:

- Problems with functionalities, security, etc.
- Ways to rectify those problems
- User expectations …

Target #1: HHS Breach Reports

Example 1

1. The covered entity (CE) experienced a cyberattack that resulted in unauthorized access to several of its websites.
2. The hackers were then able to access databases containing the protected health information (PHI) of 2,860 individuals due to a website coding error.
3. The compromised PHI included clinical, demographic, and financial information.
4. The CE provided breach notification to HHS, affected individuals, and the media.
5. Following the breach, the CE modified the coding error, moved all databases containing PHI to its internal secure network, implemented a new software patch management policy, and activated new logging and monitoring systems.
6. OCR obtained documented assurances that the CE implemented the corrective action steps listed above.

Target #2: App Reviews

Example 2

★☆☆☆☆ username1, 06/25/2014

Wifi?

I'm trying to sign up and on the part where you write your username, I press done after I type it and it brings up a message saying to check my connection. ...I've checked my connection and I've re-downloaded the app. It won't work!! Please fix it.

Research Questions

RQ_{event}	• How can we effectively extract targeted <u>events</u> from text?
RQ_{pair}	• How can we effectively extract targeted <u>event pairs</u> from text?
RQ_{story}	• How can we effectively extract targeted <u>stories</u> from text?

Ember: Extracting Targeted Events (RQ1)

Background: Structures of HHS Breach Reports

- Breach description
  - "Two unencrypted laptops were stolen from the CE's premises …"
- PHI detail
  - "The PHI involved in this breach included names, birth dates …"
- Notification
  - "The CE notified HHS, the affected individuals, and media."
- Corrective events
  - "The CE installed bars on the windows …"
- Others
  - "The OCR obtained assurances that the CE implemented the corrective action steps listed above."

Çorba: A Previous Study [Guo et al., 2020]

Norms provide a natural formal representation for security and privacy requirements

Type: c: Commitment
Subject: Covered Entity
Object: Patients
Antecedent: TRUE (at all times)
Consequent: train employee on data loss, data protection

Type: p: Prohibition
Subject: Employee
Object: Covered Entity
Antecedent: portable devices contain PHI
Consequent: lose portable devices

RQ: How can we design a crowdsourcing task to extract security requirements from regulations and breach reports as norms, and what factors affect the performance of crowd workers for this task?

Çorba: Method

- Multiple iterations to refine survey questions
  - Consequent: What actions should be (should've been) done?
  - Subject: Who should take the action?
  - Antecedent: When (in what circumstances) should the action be taken?
  - Object: Whom does (would) a breach affect?
  - Other questions, e.g., which sentences include the information?
- Evaluation (of responses)
  - Format of the question?
  - Order of the question?
  - Setup of the crowdsourcing project?
- Collection (of norms)

Çorba: Discussions

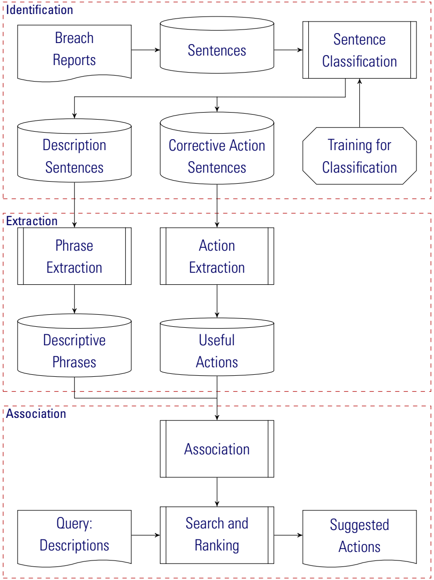
- Merits:
  - \+ Scalable norm extraction from textual artifacts
  - \+ Structured reports elicit high-quality responses
- Limitations:
  - — Results cannot be directly leveraged for automated methods
  - — Relations between norms and breach types

Ember: Research Questions

RQ_{event} How can we effectively extract informative events that provide insights to similar entities from breach reports?

$RQ_{suggest}$ How can we suggest actions to potential covered entities based on breach descriptions and common practices?

Ember: Method



Targeted HHS breach reports:

Table 1: Number of reports by length.

Number of Sentences	Count of Reports
5	628
6	541
7	395
8	177
9	89
10	43
Total	1 873

Ember: Identification of Informative Sentences

- Training set:
  - Crowdsourcing
  - **Descriptive**, **Corrective**, Neither
  - Cohen's Kappa = 0.693
- Baseline:
  - Heuristics for **PHI detail**, **Notification**, **OCR**
  - Breach reports begin with **Descriptive**
  - Others are **Corrective** sentences
- Sentence Classification:
  - Universal Sentence Encoder (USE) [Cer et al., 2018] + SVM
  - Fine-tuned BERT [Devlin et al., 2019]

Table 2: Numbers of sentences with different labels in the training set.

Sentence Type	Count
Breach Description	534
Corrective Event Sentences	448
Neither	518
Total	1 500

Ember: Extraction of Informative Phrases

Techniques:

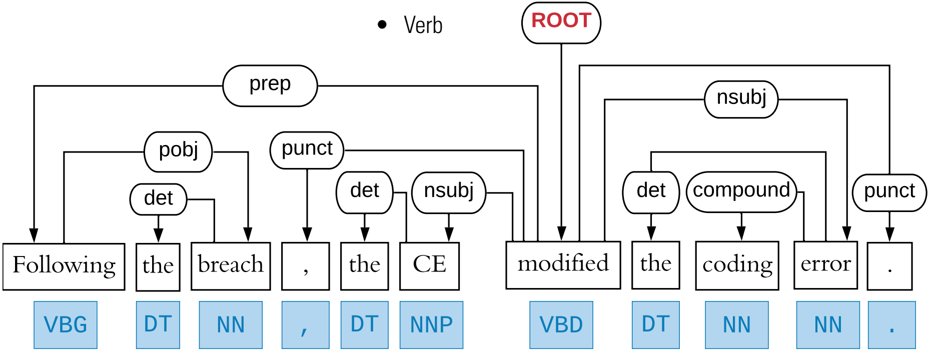
- Part-of-speech (POS) tagging
- Dependency parsing (DP)

Descriptive: POS tagging

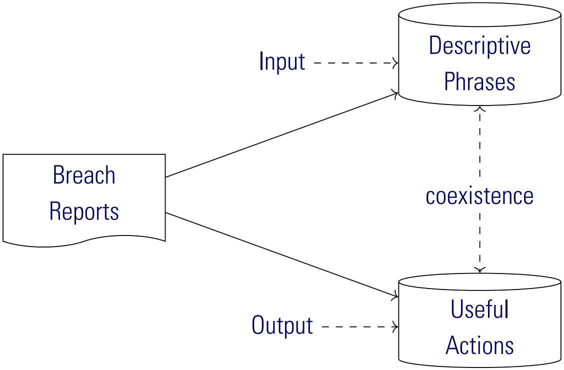
- Adjective
- Adverb
- Noun
- Verb

Corrective: DP

- Find verbs
- Find their children



Ember: Association of Descriptive Phrases and Useful Actions



- Counting actions with weights
  - More weight if report contains input phrases
- Duplicate verb phrases
  - USE + cosine similarity
- Similar descriptive phrases:
  - USE + cosine similarity

Action Suggestion Example

Action suggestion tool: https://hguo5.github.io/ActionSuggestion/

Ember: Discussions

- Merits:
  - + Automated action extraction from breach reports
  - + First tool for action suggestion
- Limitations:
  - — Limited training set for classification
  - — Association, not causal relation

Event Inference for Action Suggestion

- Event inference for action suggestion [Guo et al., 2018]

- Story Cloze Test [Mostafazadeh et al., 2016]:
  - Given a sequence of events, can a model automatically infer the probable following events?

- RQ: Given a sequence of events (breach or actions), can a model automatically suggest possible follow-up actions?

- A sequence prediction problem

Inference: Example Results

Input:

Unencrypted laptops were stolen

Actual$_1$
- The CE reported the theft to the law enforcement

Actual$_2$
- The CE recovered the laptops

Actual$_3$
- The CE implemented new policies

Actual$_4$
- The CE placed an accounting of disclosures

Pred$_1$
- The CE filed a police report to recover the stolen item

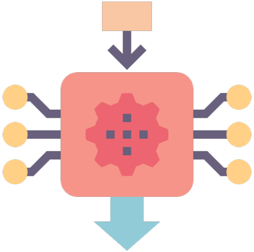
Pred$_2$
- The CE replaced its building alarm

Pred$_3$
- The CE revised its existing policies

Pred$_4$
- The CE mandated encryption for all mobile devices

Inference: Discussions

- Merits:
  - ＋ Action suggestion based on event inference
  - ＋ Toward simpler and more structured breach reporting
- Limitations:
  - — Limited training set for inference
  - — Not full inference based on causal relations



Caspar: Extracting Targeted Event Pairs (RQ2)

Motivation

- App reviews:
  - De facto deployment reports
- Action-problem pairs:
  - User action event
  - App problem event

Example 4a

★☆☆☆☆ username2, 07/14/2014

App crashing

App keeps crashing when I go and log my food. Not all the time but at least a crashing session a day.

Example 4b

★☆☆☆☆ username3, 09/12/2014

App full of bugs

The app crashes and freezes constantly. The only reason I still own a fitbit is the website.

Caspar: Event Extraction, Classification, and Order

- Event extraction
  - Dependency parsing
- Event classification:
  - USE + SVM
  - Manual labeling for training set
- Event ordering:
  - Heuristics

Table 3: Heuristics for event ordering.

Sentence Structure	Event Order
e_1, *before / until / then* e_2	$e_1 \rightarrow e_2$
e_1, *after / whenever / every time / as soon as* e_2	$e_2 \rightarrow e_1$
e_1, *when* e_2	$e_1 \rightarrow e_2$, if verb of e_1 is VBG $e_2 \rightarrow e_1$, otherwise

Extracted Action-Problem Pairs

Table 4: Extracted event pairs for the Weather Channel.

User Action		App problem
(after) I upgraded to iPhone 6	$\rightarrow$	this app doesn't work
(as soon as) I open app	$\rightarrow$	takes me automatically to an ad
You need to uninstall app	$\rightarrow$	(before) location services stops
(every time) I try to pull up weather	$\rightarrow$	I get "no data"
(whenever) I press play	$\rightarrow$	it always is blotchy
(when) I have full bars	$\rightarrow$	Always shows up not available
I updated my app	$\rightarrow$	(then) it deleted itself

Problem: Given a user action, what app problems follow?

- Event follow-up classification
  - Given a User Action and an App Problem, $\langle e_u, e_a \rangle$, is e_a a valid *follow-up event* to e_u or a *random event*?
  - USE + SVM
  - biLSTM network + Word Embedding
- Negative sampling
  - Use random examples as negative ones
  - What about similar events?
- Inference: rank possible follow-up events by probability

Caspar: Discussions

- Merits:
  - ＋ Informative: action-problem pairs
  - ＋ Predictive: event inference
- Limitations:
  - — Key phrases limit the dataset
  - — An action-problem pair may not be the whole story
  - — Event inference needs improving

Scheture: Extracting Targeted Stories (RQ3)

Motivation

- Users tell different stories
- Different stories serve different goals

Example 2, again

★☆☆☆☆ username1, 06/25/2014

Wifi?

I'm trying to sign up$_{\text{intention}}$ and on the part where you write your username, I press done after I type it$_{\text{action}}$ and it brings up a message saying to check my connection$_{\text{behavior}}$. ...I've checked my connection$_{\text{reaction}}$ and I've re-downloaded the app$_{\text{reaction}}$. It won't work$_{\text{behavior}}$!! Please fix it.

Structure:

— patterns of event types

intention ➠ action ➠ behavior ➠ reaction ➠ reaction ➠ behavior

Security and Privacy Violations

- Stories where users' expectations are not met

Example 5

★☆☆☆☆ username4, 11/28/2018

Yelpers Beware!

For 7 years, I Yelped about area restaurants, events, activities, etc. for 5 years, I was a Yelp Elite, which meant I got invited to special events for free to do cool stuff. I amassed close to 300 reviews, innumerable followers & "friends." Beware, if you write even the vaguest negative word in your review and get harassed by a biz owner, Yelp turns a blind eye. Biz owners have stalked me, threatened me, threatened to sue me, sent me hateful msgs, and the like. And note that of the near-300 reviews, 75% receive 4 or 5 stars. It's all cool if you are into PR writing & edit out any gory bad details. Good luck. The whole site is a sham.

Event Types

(I) INTENTION:

— "I wanted to update a status on Facebook"

(A) ACTION:

— "I typed it all out"

(B) BEHAVIOR:

— "It took at least 5 minutes for it to show"

(R) REACTION:

— "I deleted it and use safari instead"

(C) CONTEXT:

— "I have strong wifi signal & good service and 4 bars of service"

Scheture: Sequencing of Stories

Example 5

★☆☆☆☆ username4, 06/10/2014

HATING SO MUCH LATELY!

I HATE how in iphones you can not zoom in to record a video$_{\text{Behavior}}$. If you zoom in and try to record$_{\text{Action}}$ it goes back to normal$_{\text{Behavior}}$. How ANNOYING! I also HATE how when someone sends me a conversation$_{\text{Action}}$ my music will stop playing$_{\text{Behavior}}$ because I opened what they sent me$_{\text{Action}}$. It's not a snap necessarily$_{\text{Context}}$ it's a simple conversation$_{\text{Context}}$. Also my snapchat sometimes says like memory full$_{\text{Behavior}}$ when I try to take or record a snapchat$_{\text{Action}}$. It's so ANNOYING.

Assumptions for structure analysis:

- NONTARGET does not contribute
- Context can appear anywhere
- Adjacent events of the same type can be grouped together

Table 5: Story pattern in the examples.

Story	Review	Pattern
s_1	Example 2	I, A, B, R+, B
s_2	Example 5	B, A, B
s_3	Example 5	A+, B
s_4	Example 5	A, B

Scheture: Sequencing of Stories

- Input: Two events (e_1, e_2)
- Output: $e_1 \rightarrow e_2$, $e_2 \rightarrow e_1$, separate
- Event Relations:
  - Heuristics
  - Three-class classification
    - Word Vectors (Word2Vec, GloVe [Pennington et al., 2014])
    - Universal Sentence Encoder
    - SVM, MLP, biLSTM

Table 6: Heuristics for event relations.

Event Order	Sentence Structure
$e_1 \rightarrow e_2$	e_1, *before / until / then* e_2 e_1 [SEP] *And then* e_2
$e_2 \rightarrow e_1$	e_1, *after / when / whenever / every time / as soon as* e_2 e_1, *if / because* e_2
Separate	e_1 [SEP] *Also / Additionally* e_2

Scheture: Collection of Targeted Stories

- **Simple Reviews**
  - Reviews with one target event
- **Simple Stories**
  - Stories with one target event
- Collect by pattern matching
- Common patterns in **Complex Stories**
  - Generalized Sequential Pattern (GSP)

Results: Events and Stories

- Event type classification accuracy: 74.1% (SVM)
- Types of reviews:
  - 17.63%: Without target events
  - 22.44%: Simple Reviews
  - 59.94%: Complex Reviews
- Training for event relation classification:
  - 1 005 166 event pairs from heuristics (32.4%)
  - Randomly sampled 60 000 pairs (20 000 for each type)
  - 90% for training and 10% for testing
- Event relation classification accuracy: 79.7% ($\text{BERT}_{\text{base}}$)

Results: Collection of Targeted Stories

- Intention (I), Action (A), Behavior (B), and Reaction (R) events only
- 2 500 580 stories from Complex Reviews:
  - Context only: 269 409 (10.8%)
  - **Simple Stories**: 1 558 156 (62.3%)
  - **Complex Stories**: 673 015 (26.9%)

Table 7: Common story structures.

Simple Stories		Complex Stories (freq $>$ 1%)					
Length 1		Length 2		Length 3		Length 4	
B	855 630 (54.9%)	AB	176 661 (26.25%)	BAB	39 291 (5.84%)	ABAB	8 869 (1.32%)
B+	365 361 (23.4%)	BR	85 807 (12.75%)	BRB	19 794 (2.94%)		
R	152 259 (9.77%)	BA	60 310 (8.96%)	ABR	13 030 (1.94%)		
A	88 178 (5.66%)	RB	56 928 (8.46%)	ABA	9 431 (1.40%)		
I	55 613 (3.57%)	AB+	52 817 (7.85%)	BAB+	7 783 (1.16%)		
R+	25 592 (1.64%)	IB	34 629 (5.15%)				
A+	12 747 (0.82%)	B+R	20 414 (3.03%)				
I+	2 776 (0.18%)	BI	16 091 (2.39%)				
		B+A	12 858 (1.91%)				
		AR	12 486 (1.86%)				
		RB+	9 943 (1.48%)				
		A+B	9 815 (1.46%)				
		IB+	8 424 (1.25%)				
		RA	7 793 (1.16%)				
		R+B	7 249 (1.08%)				
		BR+	7 075 (1.05%)				

Results: Common Patterns

Table 8: Frequent substructures (freq $>$ 1%) in Complex Stories.

Length 1		Length 2		Length 3		Length 4	
B	629 562 (93.54%)	AB	294 096 (43.70%)	BAB	67 178 (9.98%)	ABAB	14 760 (2.19%)
A	422 417 (62.76%)	BR	161 000 (23.92%)	BRB	34 883 (5.18%)	ABRB	7 162 (1.06%)
R	285 226 (42.38%)	BA	157 842 (23.45%)	ABA	30 222 (4.49%)	BABR	7 025 (1.04%)
B+	193 518 (28.75%)	RB	115 699 (17.19%)	ABR	28 600 (4.25%)	BABA	6 743 (1.00%)
I	117 656 (17.48%)	AB+	85 970 (12.77%)	B+AB	14 836 (2.20%)		
A+	41 255 (6.13%)	IB	59 579 (8.85%)	BAB+	13 618 (2.02%)		
R+	37 069 (5.51%)	B+R	44 761 (6.65%)	RBR	13 261 (1.97%)		
		B+A	39 033 (5.80%)	BAR	12 690 (1.89%)		
		BI	37 440 (5.56%)	ARB	10 759 (1.60%)		
		AR	37 371 (5.55%)	BIB	10 595 (1.57%)		
		A+B	28 471 (4.23%)	RAB	10 536 (1.57%)		
		RA	28 092 (4.17%)	BRA	9 424 (1.40%)		
		RB+	21 953 (3.26%)	AB+R	8 068 (1.20%)		
		R+B	16 642 (2.47%)	B+RB	8 050 (1.20%)		
		IA	15 670 (2.33%)	RBA	7 719 (1.15%)		
		IB+	14 580 (2.17%)	AB+A	7 429 (1.10%)		
		BR+	14 382 (2.14%)				
		AI	13 530 (2.01%)				
		IR	13 178 (1.96%)				
		BA+	11 381 (1.69%)				
		A+B+	9 182 (1.36%)				
		B+I	8 902 (1.32%)				
		RI	7 525 (1.12%)				

Results: Extracted Stories

B+

[B]	*This new format is so awful*
[B]	*Half the time it "can not get weather data"*
[N]	*(When) it does*
[B]	*it is slow to load and difficult to navigate*

AB

[A]	*(when) I'm typing to another person*
[C]	*& they are there*
[B]	*The yellow button doesn't always turn blue*
[N]	*FIX IT SNAPCHAT!*

ABRB

[N]	*I love Pandora*
[A]	*I just started listening to Pandora*
[B]	*(But often times) I'm unable to skip songs*
[R]	*I've tried quitting and reopening…*
[B]	*None of which work/help!!*
[N]	*What's up with this?*

IABR

[I]	*I want to be able to delete saved chats!!!*
[A]	*(Because if) I accidentally tap a message*
[B]	*(then) it becomes bolded font and saves*
[R]	*(yet) I can't unsave it!*
[N]	*FIX IT!!!*

Manual Verification: Are stories with patterns more helpful than random stories?

Table 9: Average helpfulness scores of different stories toward different goals (p_s denotes p-value against simple problem stories; p_r denotes p-value against random stories).

Goal	Simple Problem Stories	Random Stories	Pattern	Score	p_s	p_r
App Problem	3.578	3.435	A+B+	$\boxed{4.163}$	0.003	0.000
			C+B+	$\boxed{4.118}$	0.009	0.000
			B+R+	$\boxed{4.136}$	0.005	0.000
			I+A+	3.900	-	-
User Retention	1.689	1.825	A+B+	1.596	-	-
			C+B+	1.735	-	-
			B+R+	$\boxed{2.652}$	0.001	0.005
			I+A+	1.617	-	-
User Expectation	3.467	3.275	A+B+	3.125	-	-
			C+B+	3.039	-	-
			B+R+	2.288	-	-
			I+A+	$\boxed{4.133}$	-	0.000

Scheture: Discussions

- Merits:
  - + Systematic way to search for stories
  - + More event types
  - + Event sequencing
- Limitations:
  - — Are the targeted event types enough?
  - — Is parser-based extraction reliable enough?
  - — Are the classifications good enough?

Unexpected Information Access

Unexpected Information Access (UIA)

Example - Without Consent

AirBeam Video Surveillance App

…with this app, i can spy on my family without them knowing it! it's such an awesome app!

Example - Victim is not comfortable

HER Lesbian Dating App

I had someone cyberstalking and harassing me. Multiple attempts in every way shape and form were made to contact app-name to block and ban the stalker's account due to a concern for my well- being …

Identifying UIA-Enabling Apps

How convincing?

Very *"This app is high key creepy. When I'm with my dad on his days my mom even mentions how she knew everything I was doing and it even made my dad creeped out. I don't want my mom stalking me."*

So-so *"This app is perfect for stalking people."*

Maybe *"May work well to spy on the kids by 'accidentally' leaving iPhone in a secret place."*

How severe?

Very *"This app has truthfully ruined my teenage years all because my mother now has a way of tracking me down 24/7. I couldn't do the normal teenage things because I was being stalked all day."*

So-so *"My boyfriend sees my location which is bit creepy, but I realize it's nice to track each other for safety."*

Conclusion

Conclusion

- We targeted text related to software development
- We investigated:
  - Extracting informative events
  - Extracting informative event pairs
  - Extracting informative stories
- Future work:
  - More reliable extraction from low-quality text
  - Pre-defined event types
  - Deeper understanding of event relations
  - How does story understanding help?

Thank you! Questions?

Email: hguo5@ncsu.edu

URL: https://hguo5.github.io/phddefense/

Appendix

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175:1–7, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Hui Guo, Özgür Kafalı, and Munindar P. Singh. Extraction of natural language requirements from breach reports using event inference. In *International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, pages 22–28, Banff, AB, Canada, August 2018. IEEE Press.

Hui Guo, Özgür Kafalı, Anne-Liz Jeukeng, Laurie Williams, and Munindar P. Singh. Çorba: Crowdsourcing to obtain requirements from regulations and breaches. *Empirical Software Engineering*, 25(1):532–561, January 2020.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.