

Informative Events and Stories in Text Related to Software Development

Final Oral Exam

Hui Guo

Under the guidance of Dr. Munindar P. Singh

May 11th, 2021

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

Email: hguo5@ncsu.edu

NC STATE UNIVERSITY

1. Introduction
2. Lesbre: Extracting Targeted Events (RQ1)
3. Caspar: Extracting Targeted Event Pairs (RQ2)
4. Scheture: Extracting Targeted Stories (RQ3)
5. Conclusion

Introduction



Natural language text in software engineering

Stories: prevalent in NL text

Rich information about:

- Where the problems are
- How to rectify those problems
- What is needed ...

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.

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.

RQ_{event}

- How can we effectively extract targeted events from text?

RQ_{pair}

- How can we effectively extract targeted event pairs from text?

RQ_{story}

- How can we effectively extract targeted stories from text?

Lesbre: 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.”



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?


- 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)

**AND THE
SURVEY
SAYS...**




Çorba Results: 60 Unique Norms from 38 Breach Reports

ID	Task	Response
323	What	The portable drive should have been better safeguarded, including using data encryption
	Who	The pharmacy resident
	When	When handling patients' data it should always be encrypted and handled with the utmost concern
	Whom	Arnold Palmer Hospital
344	What	Provide ample training to residents
	Who	Arnold Palmer Hospital
	When	When training for residents is necessary
	Whom	HHS and patients



c(employee, CE, portable devices contain PHI, safeguard portable devices)



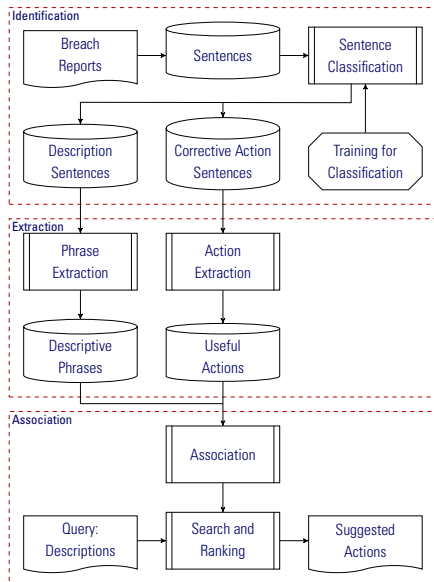
c(CE, patients, TRUE, train employees on data loss and data protection)

- 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



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?



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

- 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

Lesbre: Extraction of Informative Phrases

Techniques:

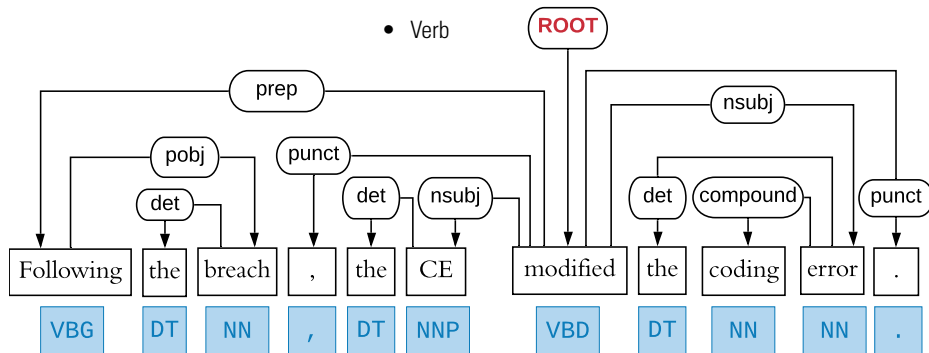
- Part-of-speech (POS) tagging
- Dependency parsing (DP)

Descriptive: POS tagging

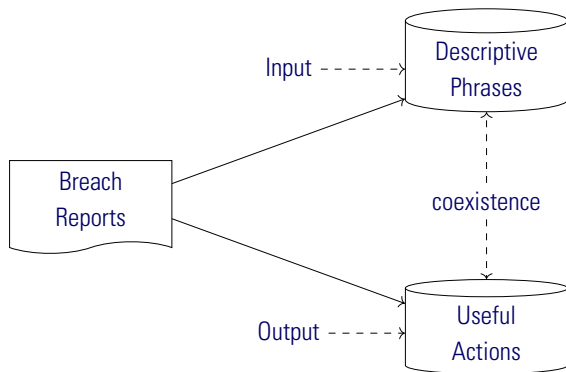
- Adjective
- Adverb
- Noun
- Verb

Corrective: DP

- Find verbs
- Find their children



Lesbre: Association of Descriptive Phrases and Useful Actions



- Counting actions with weights
 - More weight if report contains input phrases
- Clustering the actions
 - USE + cosine similarity
 - DBScan + K-means
- Similar descriptive phrases:
 - USE + cosine similarity

Table 3: Accuracy of classification.

Classifier	Accuracy
USE+SVM	94.0%
Fine-tuned BERT	94.7%
Baseline	86.2%

Table 4: Distribution of sentence types.

Sentence Type	Count
Breach Description	4 176 (35.1%)
Corrective Event Sentences	3 911 (32.8%)
Neither	3 819 (32.1%)
Total	11 906

- 26 092 descriptive phrases
- 4 770 clusters of useful actions
- Action suggestion tool:

<https://hgao5.github.io/ActionSuggestion/>

Action Suggestion Example

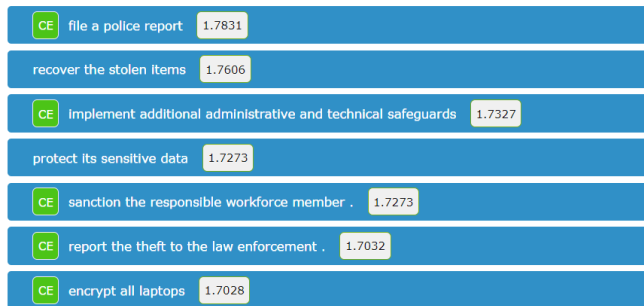
Likely key phrases to describe the breach:



Search keys for the breach:



Suggested Actions:



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



- Story Cloze Test [Mostafazadeh et al., 2016] and ROCStories [Mostafazadeh et al., 2017]
- Given a sequence of events, can a model automatically infer the probable following events?

Example 3a

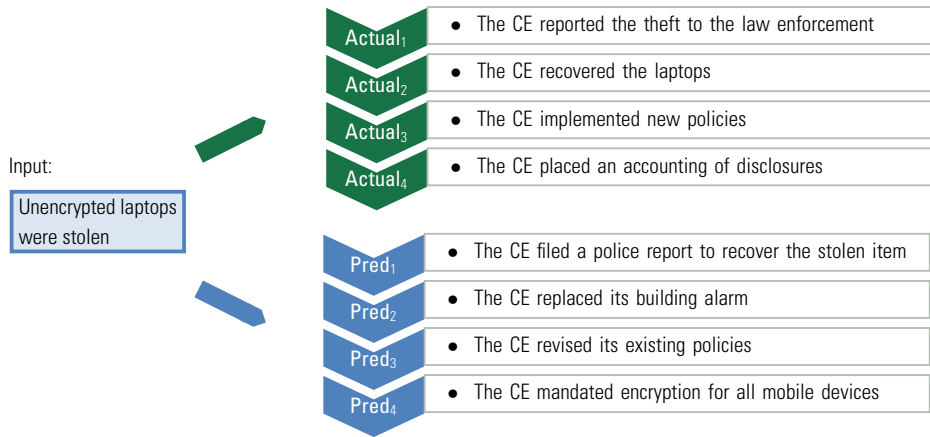
1. Tonight I played 3 games of online speed chess with Jim.
2. During the first game, the board froze.
3. Jim was able to checkmate me.
4. I signed off and went back on.
5. I won the next game.

Example 3b

1. (D) Two laptop computers with questionable encryption were stolen from the CE's premises.
2. (A1) The CE reported the theft to the law enforcement.
3. (A2) The CE worked with the local police to recover the laptops.
4. (A3) The CE developed and implemented new policies and procedures to comply with the Security Rule.
5. (A4) The CE placed an accounting of disclosures in the medical records of all affected individuals.

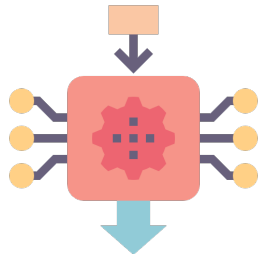
- Event inference for action suggestion [Guo et al., 2018]
- RQ: Given a sequence of events (breach or actions), can a model automatically suggest possible follow-up actions?
- A sequence prediction problem:
 - Average Word2Vec, Paragraph Vector (Doc2Vec) [Le and Mikolov, 2014; Mikolov et al., 2013]
 - LSTM network [Hochreiter and Schmidhuber, 1997]
 - Multiple models for multiple follow up events

Inference: Example Results



- Manual verification: 60% plausible, 35% matching actual events

- 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)

- App reviews:
 - De facto deployment reports
- App-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

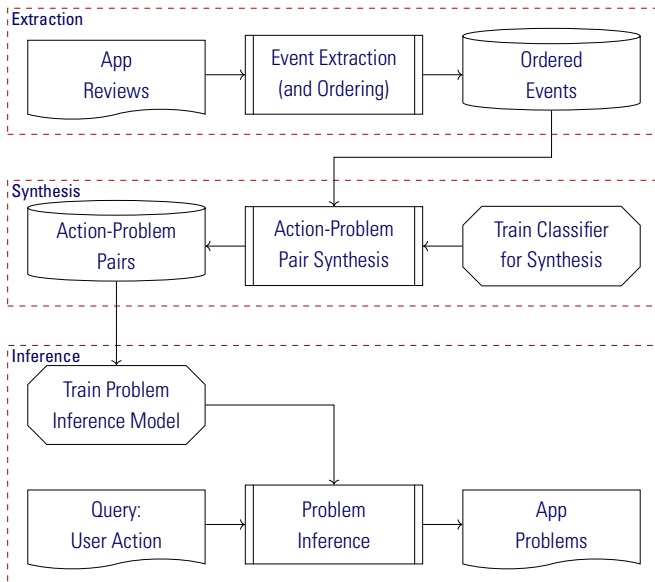
App full of bugs

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

RQ_{pair} How effectively can we extract action-problem pairs from app reviews?

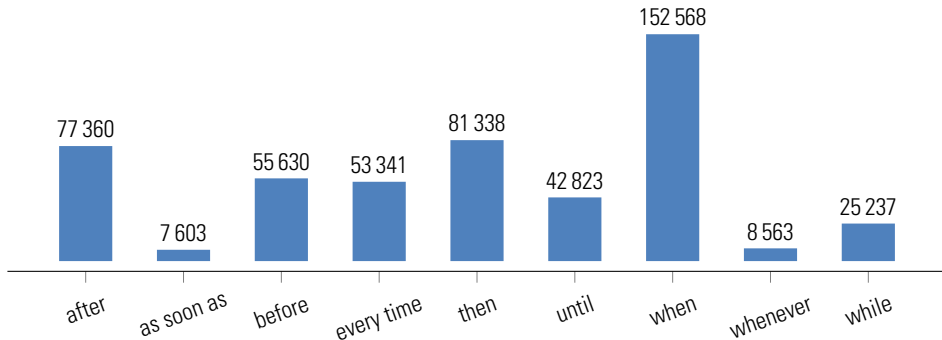
$RQ_{\text{infer-pair}}$ How effectively can an event inference model infer app problems in response to a user action?

Caspar: Overview



Dataset

- Apple App Store, 58 674 198 reviews, 151 apps
- 1 220 003 one-star reviews
- 393 755 reviews with key phrases



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

Table 5: Distribution of the manually labeled dataset.

Event type	Count
User Action	401
App Problem	383
Neither	602
Total	1 386

Table 6: Heuristics for event ordering.

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

Table 7: Extracted event pairs for the Weather Channel.

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

Results:

- Accuracy: 82.0%
- 85 099 action-problem pairs

More info at:

<https://hgao5.github.io/Caspar/>

Table 8: Man vs. Caspar.

		All reviews		Reviews w/ key	
		Human		Human	
		ID-ed	Not ID-ed	ID-ed	Not ID-ed
Caspar	ID-ed	13/200	1/200	13/63	1/63
	Not ID-ed	25/200	161/200	16/63	33/63

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

- Accuracy: 82.8%
- Manual verification

User Action: I try to scroll thru cities

Ground truth: it hesitates

Inferred App Problems:

Relevant

- a_1 it says there is an error
- a_2 it loads for what seems like forever
- a_3 it tells me the info for my area is not available
- a_4 the app crashes
- a_8 it reset my home location

Conflicting judgments

- a_6 it rarely retrieves the latest weather without me having to refresh
- a_9 it goes to a login screen that does not work

Irrelevant

- a_5 the radar never moves , it just disappears
- a_7 I rely heavily on it & for the past month , it says temporarily unavailable
- a_{10} Radar map is buggy – weather activity stalls , appears , then disappears

- 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



Schettre: 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_{intention} and on the part where you write your username, I press done after I type it_{action} and it brings up a message saying to check my connection_{behavior}. ...I've checked my connection_{reaction} and I've re-downloaded the app_{reaction}. It won't work_{behavior}!! Please fix it.

Structure:

— patterns of event types

intention ➡ action ➡ behavior ➡

reaction ➡ reaction ➡ behavior

(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”

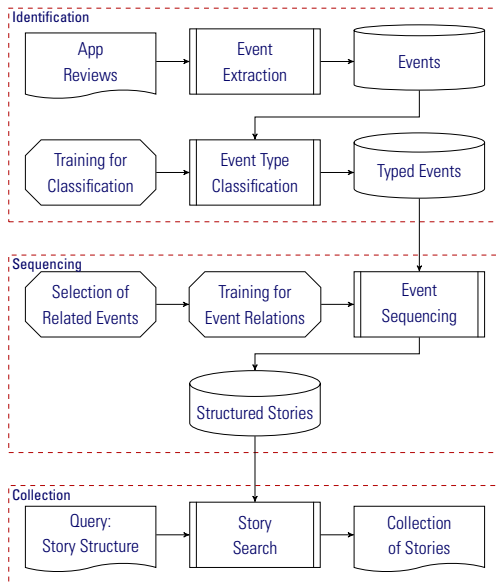


RQ_{event} How effectively can we extract events and determine their types in app reviews?

RQ_{relate} How effectively can we identify relations between events, so that we can order and combine them into stories?

RQ_{collect} What kind of story structures and substructures are the most common in app reviews?

Schetchure: Overview



Data collection: Crowdsourcing

Table 9: Distribution of event types.

Event Type	Seed Dataset	Final Dataset
Intention	23 (7.67%)	263 (8.77%)
Action	32 (10.67%)	422 (14.07%)
Behavior	101 (33.67%)	741 (24.70%)
Reaction	37 (12.33%)	531 (17.70%)
Context	58 (19.33%)	472 (15.73%)
Nontarget	49 (16.33%)	571 (19.03%)
Total	300	3 000

Event Identification:

- Sentence Embedding
 - Universal Sentence Encoder (USE)
 - Event vs. Context
- Text classifiers (six-class):
 - SVM
 - MLP
 - KNN
 - DT

Example 5

★★★★☆ username4, 06/10/2014
HATING SO MUCH LATELY!
I HATE how in iphones you can not zoom in to record a video_{Behavior}. If you zoom in and try to record_{Action} it goes back to normal_{Behavior}. How ANNOYING! I also HATE how when someone sends me a conversation_{Action} my music will stop playing_{Behavior} because I opened what they sent me_{Action}. It's not a snap necessarily_{Context} it's a simple conversation_{Context}. Also my snapchat sometimes says like memory full_{Behavior} when I try to take or record a snapchat_{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 10: 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

Schecture: 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 11: Heuristics for event relations.

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

Schecture: 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: Identification of Target Events

9 305 505 event phrases from 2 118 942 reviews

- 373 470 Without target events (17.63%)
- 475 445 Simple Reviews (22.44%)
- 1 270 027 Complex Reviews (59.94%)

Table 12: Event type classification.

Model	Accuracy	Precision	Recall	F-1 Score
KNN _{event}	0.661	0.922	0.951	0.936
SVM _{event}	0.741	0.956	0.932	0.944
DT _{event}	0.539	0.896	0.906	0.901
MLP _{event}	0.717	0.953	0.930	0.942
KNN _{context}	0.584	0.888	0.954	0.920
SVM _{context}	0.718	0.955	0.914	0.934
DT _{context}	0.529	0.894	0.910	0.902
MLP _{context}	0.720	0.949	0.932	0.941

Table 13: Distribution of event types.

Event Type	Event Count	Simple Reviews
Intention	203 053 (2.18%)	16 256 (3.42%)
Action	658 931 (7.08%)	31 377 (6.60%)
Behavior	3 065 360 (32.94%)	334 549 (70.37%)
Reaction	591 624 (6.36%)	35 507 (7.47%)
Context	1 201 579 (12.91%)	57 756 (12.15%)
Nontarget	3 584 958 (38.53%)	—
Total	9 305 505	475 445

Results: Sequencing Stories

- 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
- 2 500 580 stories in Complex Reviews

Table 14: Event relation classification.

Model	Accuracy
SVM _{GloVe}	0.737
SVM _{Word2Vec}	0.728
SVM _{USE}	0.752
MLP _{GloVe}	0.727
MLP _{Word2Vec}	0.718
MLP _{USE}	0.736
LSTM _{GloVe}	0.722
LSTM _{Word2Vec}	0.714
BERT _{base}	0.797

Results: Collection of Targeted Stories

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

Table 15: 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%)				

Table 16: 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*

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?*

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!*

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 17: 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+	4.163	0.003	0.000
			C+B+	4.118	0.009	0.000
			B+R+	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+	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+	4.133	-	0.000

- 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?



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.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, November 1997.
- Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, pages 1188–1196, Beijing, China, 2014. Omnipress.

- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS*, pages 3111–3119, Lake Tahoe, Nevada, December 2013. Neural Information Processing Systems Foundation.
- 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.

- Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain, 2017. Association for Computational Linguistic.
- 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.