4-27-2013

Script-Based Story Matching for Cyberbullying Prevention

Jamie Macbeth
Massachusetts Institute of Technology, jmacbeth@smith.edu

Hanna Adeyema
Massachusetts Institute of Technology

Henry Lieberman
MIT Media Lab

Christopher Fry
MIT Media Lab

Script-Based Story Matching for Cyberbullying Prevention

Jamie Macbeth
MIT Humans and Automation Laboratory
77 Massachusetts Ave.
Building 33-407
Cambridge, MA 02139 USA
jmacbeth@mit.edu

Hanna Adeyema
MIT
21 Parkvale Ave. #8
Allston, MA 02134 USA
adeyema@mit.edu

Henry Lieberman
MIT Media Laboratory
20 Ames St., 320 F
Cambridge, MA 02139 USA
lieber@media.mit.edu

Christopher Fry
MIT Media Laboratory
20 Ames St., E15-358
Cambridge, MA 02139 USA
cfry@media.mit.edu

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract
While the Internet and social media help keep today's youth better connected to their friends, family, and community, the same media are also the form of expression for an array of harmful social behaviors, such as cyberbullying and cyber-harassment. In this paper we present work in progress to develop intelligent interfaces to social media that use commonsense knowledge bases and automated narrative analyses of text communications between users to trigger selective interventions and prevent negative outcomes. While other approaches seek merely to classify the overall topic of the text, we try to match stories to finer-grained "scripts" that represent stereotypical events and actions. For example, many bullying stories can be matched to a "revenge" script that describes trying to harm someone who has harmed you. These tools have been implemented in an initial prototype system and tested on a database of real stories of cyberbullying collected on MTV's "A Thin Line" Web site.

Author Keywords
Commonsense reasoning; affective computing

ACM Classification Keywords
H.1.2 [**Models and Principles**]: Human information processing; H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: Natural Language

Figure 1: Over The Line? Web site, part of MTV's A Thin Line campaign.

Introduction

The near-ubiquity of social networks and media have enabled young people in our society to be more connected to friends, family and community while enhancing and creating new and interesting forms of positive social interaction and learning. However, these technologies have also enabled a set of negative and harmful social behaviors; cyberbullying and cyber-harassment have been identified as a major risk factor facing young people and even many adults. Awareness of cyberbullying has increased rapidly over the past decade and studies indicate that about half of teens and adolescents have experienced cyberbullying in some form at least once [5]. Other studies indicate how peer harassment experiences contribute to depression, decreased self-worth, hopelessness, and suicidal ideation among young people [4].

Campaigns to prevent cyberbullying typically use videos and print media to provide general education on the problem to users. Some of the videos are interviews with celebrities describing their previous experience of being bullied or cyberbullied. However the social media technologies make it possible to deliver useful advice to end-users that is contextualized to the specific situations they are in at the time that they arise. For example, reflective user interfaces can allow potential victims see stories from other youth in similar situations. They could provide perpetrators with educational material and provide third parties with links to information on how to provide emotional support to victims. However, the complexity, creativity and contextual specificity of bullying comments and situations make them extremely difficult to detect and interpret using the standard natural language processing toolset.

In this work-in-progress paper we motivate the need for better tools to develop intelligent interfaces based on story understanding of natural language descriptions at various levels of generality. We propose that automated, intelligent story matching will enhance and improve the accuracy of all tools and human interfaces related to cyberbullying prevention. We have composed a story matcher tool and interface that use standard natural language processing tools for parsing English sentences in combination with a commonsense knowledge base used to relate words describing actions, objects and locations of actions, implements used in actions, and actors roles. The system calculates a similarity metric between stories by aggregating similarity metrics of the words appearing as concepts in the commonsense database. We also consider general challenges of story understanding for cyberbullying prevention.

Challenges of Narrative Comprehension

The shifting and increasing capabilities of mobile devices and Web technologies have made it difficult to determine potential negative consequences of sending or posting material about oneself or others in digital form. The Web site "A Thin Line" (www.athinline.org) exists as part of a MTV campaign to prevent digital abuse by empowering people to understand and identify it in its many forms.

A special section of the site called Over the Line? is devoted to stories of user behavior in the social media space. It allows users to post their stories so that others can vote on whether the described behavior is socially acceptable or unacceptable by rating it "under the line" or "over the line" respectively. Stories are limited to 250 characters and may be anonymous, or the user may optionally give their first name, age and sex.

The site has collected thousands of stories from young people related to cyberbullying, and commonsense knowledge and reasoning have been valuable in automatically detecting cyberbullying messages [2]. However many instances of cyberbullying are complex stories involving multiple actors and chains of actions and events. For example, one of the most common situations is someone posting nude pictures of their ex-girlfriend or ex-boyfriend as an act of revenge. Even with variations, the "revenge script" is easily identifiable by humans. A victim first detects a negative act by former or current partner and then reacts as a perpetrator by distributing negative information about the partner on a social networking site. The following are two stories contributed by users to Over The Line?:

Revenge-1

"My ex-girlfriend accused me of doing drugs and being gay and stingy, announcing her point of view in public on my Facebook wall. What would you do? Would you delete the message and let it go, or spam her Web site!!!!?"

Revenge-2

"I thought I loved my boyfriend and I stripped for him on Skype while on vacation, but I was wrong. We had a messy break up. I moved on, and he didn't. He wanted revenge, so he sent pictures of me to my sorority and got me kicked out."

These two stories follow the same basic pattern of an ex-boyfriend or ex-girlfriend posting or sending a form of electronic media to hurt the other person as an act of revenge. This is a difficult match to detect with the standard natural language processing tools for many reasons. To name just a few:

1. Although it is implied by the reference to a boyfriend, and subsequent use of pronoun "we" in describing the the break-up, Revenge-2 never explicitly mentions the "ex-boyfriend" or "ex-girlfriend" concept.
2. Revenge-1 uses the phrase "announcing her point of view in public on my Facebook wall", which requires the reader to understand the "point of view" as the "accusation of using drugs, being gay and stingy" is what was posted on Facebook.
3. In Revenge-2, a reader will assume that the pictures sent to the sorority were intimate pictures due to the description "I stripped for him on Skype while on vacation." If the reader is unfamiliar with Skype, they can draw the conclusion that Skype is a video telecommunication system through their understanding of the stripping act. Even without knowledge or mention of the intimate Skype communication, the reader may be able to infer that a vengeful act occurred due to the mention of the photos sent causing the writer to be "kicked out" of her sorority.
4. Knowledge of human plans and goals of both the perpetrator and the victim of a cyberbullying act are also necessary to understand the act. For example in Revenge-2, a human reader understands that "...and got me kicked out." is a reference to the writer being expelled from the sorority organization mentioned in the previous sentence. As well, the human reader understands that the writer had a

plan and a goal of remaining in the sorority, and that the cyberbully sent pictures to members of the sorority as a counter-plan to hurt the writer by causing a goal failure.

5. Although Revenge-2 explicitly mentions that the cyberbully sought revenge, most revenge stories do not.

These natural language understanding subtasks—easily performed by human readers to match these stories together under the revenge "script" generalization—are difficult to accomplish with bag-of-word language models and typical information extraction methods. As an initial step towards a script matching system we have created a exploratory prototype system that allows users to search for stories matching a story template or script. The system was tested on a set of stories selected from a database on Over the Line? stories. In the system the user provides a "query sentence" which is matched against sentences in the stories. The system returns the stories in order of the greatest matching score among their constituent sentences. A full explanation of this process follows.

Sentence Dependency Parsings
Achieving a greater level of accuracy in computer-based understanding of cyberbullying stories than a bag-of-words model requires that the ordering of words and the syntactic contexts of individual words are taken into account. In our system this is performed by generating automated parsings of sentences.

The system uses the Stanford Natural Language Processing system [6] which has implementations of probabilistic natural language parsers, both highly optimized Probabilistic Context Free Grammar (PCFG) and dependency parsers, and a lexicalized PCFG parser in Java; specifically, we chose to use the dependency parser within this framework [1]. The dependency parser produces a directed, acyclic graph with words as nodes and grammatical dependency labelings on the edges. The graph is represented as an adjacency-list of the form `dependency(word,word)`, each labeled with the dependency type. Scanning the form of the parse makes it easy to find the most important elements of one sentence for comparison with another. For example, it is easy to find the main verb, subject and object of a sentence by scanning for the `ROOT`, `nsubj`, `dobj` dependency types respectively. An example a dependency parsing is shown in Figure 2.

```
nsubj(trashed-2, She-1)
root(ROOT-0, trashed-2)
dobj(trashed-2, me-3)
prep_on(trashed-2,Facebook-5)
```

Figure 2: A Stanford NLP dependency parsing of the sentence "She trashed me on Facebook." Dependency parsings allow for easy identification of key concepts in the sentence.

Sentence Matching Using Commonsense Knowledge
Two sentences are matched by iterating over each word in the source sentence and looking for a dependency of the same type in the target sentence. If a matching dependency is found in the target sentence, a score for the match is calculated using a commonsense knowledge base to relate the similarity of the two concepts. The system can then aggregate the match scores for individual dependencies into an overall match for the sentence. The strength of similarity between the two similarly-dependent concepts is calculated using the ConceptNet AnalogySpace framework [3]. ConceptNet is a semantic representation of over a million items of commonsense information collected from humans online through the Open Mind Common

Sense (OMCS) project. AnalogySpace performs reasoning over the sparse set of ConceptNet assertions using dimensionality reduction methods.

For example, in calculating a match between the stories "She trashed me on Facebook," and "He posted pictures of me on a Web site," dependency parsings of both sentences will have a `prep_on` dependency, with "Facebook" and "Web site" being the dependent words. A query is then made to AnalogySpace (using its `how_true_is` function) to determine the relative truth of the assertion "Facebook is a Web site." The score of the `prep_on` dependency match is the truth value returned by AnalogySpace. The matching continues to a restricted set of dependency types in the grammatical structure of each sentence. Knowledge relevant to Over the Line? stories is added to the commonsense database as needed.

Querying the Story Database
The current system for testing the story matching tools contains a small database of Over the Line? stories, and allows the user to enter a "query" sentence and view stories with sentences with a match score above a threshold. Dependency parses of all sentences for all stories of interest are stored in the database along with the corresponding text for the stories. A dependency parse of the query sentence is also performed. A match score is calculated between the query sentence and every sentence in each story in the database, and the sentence in each story with the greatest match score is selected as a representative of that story. The system then returns the set of stories with the highest representative match scores.

Discussion

This paper describes work-in-progress on an approach to combining NLP parsing techniques with ConceptNet's AnalogySpace to achieve a greater level of accuracy in detecting and reacting to cyberbullying comments in the social network space. A special commonsense knowledge base directly relevant to the scenarios is essential to "translate" stories about online bullying language into simpler English. Finally, a ranking system has been developed to analyze the relevance of matched stories to a base scenario.

Many instances of cyberbullying are complex stories involving multiple actors and chains of actions and events. In future work to properly detect and react to a bullying comment, we need to identify the roles of the personalities described in the comment and the actions that they have completed or are about to complete. This information is essential for reflective user interfaces and for more accurate targeted education that automatically expose participants to different types of content depending on their role in the situation. Cyberbullying is a real problem on the Internet and the Web and electronic social media. At the same time, the fact that these resources are computer-based makes accurate automated detection and prevention a possibility.

Related Work

Previous work by Dinakar et al. [2] used natural language processing methods in combination with a commonsense knowledge base specifically designed for cyberbullying detection. The commonsense knowledge space, called BullySpace, contained a set of knowledge assertions related to social and cultural stereotypes needed to comprehend typical bullying messages, messages that bullying perpetrators send to victims. The model was used to analyze real-world cyberbullying instances from a online social network called Formspring, which is popular with teenagers. They propose user interfaces to social

network applications that prevent cyberbullying by analyzing messages before delivering them in order to determine if they will be harmful. If the message is determined to be harmful, the system rejects delivery of the message and offers educational material on the possible harm the message may cause, or causes a delay in the delivery of the message, giving the sender time to consider the consequences.

Dinakar et al. use a bag-of-words supervised classifier model to analyze messages for instances of bullying. A companion works-in-progress submission [7] describes the augmentation of MTV's "Over The Line" Web site with a personalized story matcher based on these methods that classifies stories according to high-level themes. In contrast, the current paper focuses on a narrative understanding approach, presenting initial steps towards a commonsense-based story matching system with capability for deeper understanding. Schank and Abelson [8] recognized the importance of commonsense knowledge related to sequences of acts and events that frequently occur in society. Computer understanding of natural language requires this knowledge of common scenarios or scripts in order to "read between the lines" and draw conclusions when little information is given.

Acknowledgments

We thank Sila Sayan and Carlo Mannino for their ideas and contributions.

References

[1] De Marneffe, M., Maccartney, B., and Manning, C. D. Generating typed dependency parses from phrase structure parses. In *In Proc. Intl Conf. on Language Resources and Evaluation* (2006), 449–454.

[2] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst. 2*, 3 (Sept. 2012), 18:1–18:30.

[3] Havasi, C., Speer, R., and Alonso, J. ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In *Recent Advances in Natural Language Processing* (Borovets, Bulgaria, Sept. 2007).

[4] Hinduja, S., and Patchin, J. Bullying, cyberbullying, and suicide. *Archives of Suicide Research 14*, 3 (2010), 206–221.

[5] Hinduja, S., and Patchin, J. W. Stop cyberbullying before it starts. http://www.ncpc.org/resources/files/pdf/bullying/cyberbullying.pdf. Accessed: 1/6/2013.

[6] Klein, D., and Manning, C. D. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (2003), 423–430.

[7] Lieberman, H., Dinakar, K., and Jones, B. Crowdsourced ethics with personalized story matching. In *Proc. CHI* (2013).

[8] Schank, R. C., and Abelson, R. P. *Scripts, plans, goals and understanding : an inquiry into human knowledge structures*. L. Erlbaum Associates, Hillsdale, N.J., 1977.

[9] Winston, P. The strong story hypothesis and the directed perception hypothesis. In *AAAI Fall Symposium on Advances in Cognitive Systems* (2011).