The Gun Violence Database: A new task and data set for NLP

Ellie Pavlick[1] **Heng Ji**[2] **Xiaoman Pan**[2] **Chris Callison-Burch**[1]
[1]Computer and Information Science Department, University of Pennsylvania
[2]Computer Science Department, Rensselaer Polytechnic Institute

Abstract

We argue that NLP researchers are especially well-positioned to contribute to the national discussion about gun violence. Reasoning about the causes and outcomes of gun violence is typically dominated by politics and emotion, and data-driven research on the topic is stymied by a shortage of data and a lack of federal funding. However, data abounds in the form of unstructured text from news articles across the country. This is an ideal application of NLP technologies, such as relation extraction, coreference resolution, and event detection. We introduce a new and growing dataset, the Gun Violence Database, in order to facilitate the adaptation of current NLP technologies to the domain of gun violence, thus enabling better social science research on this important and under-resourced problem.

1 Introduction

The field of natural language processing often touts its mission as harnessing the information contained in human language: taking unstructured data in the form of speech and text, and transforming it into information that can be searched, categorized, and reasoned about. This is an ambitious goal, and the current state-of-the-art of language technology has made impressive strides towards understanding "who did what to whom, when, where, how, and why" (Kao and Poteet, 2007). Advances in NLP have enabled us to read news in real time (Petrović et al., 2010), identify the key players (Ruppenhofer et al., 2009), recognize the relationships between them (Riedel et al., 2013), summarize the new information (Wang et al., 2016), update central databases (Singhal, 2012), and use those databases to answer questions about the world (Berant et al., 2013).

Although these technological achievements are profound, often times we as researchers apply them to somewhat trivial settings like learning about the latest Hollywood divorces (Wijaya et al., 2015) or learning silly facts about the world, like that ⟨*white suites*, *will never go out of*, *style*⟩ (Fader et al., 2011). In this paper, we call the attention of the NLP community to one particularly good use case of our current technology, which could have profound policy implications: gun violence research.

Gun violence is an undeniable problem in the United States, but its causes are poorly understood, and attempts to reason about solutions are often marred by emotions and political bias. Research into the factors that cause and prevent gun violence is limited by the fact that data collection is expensive, and political agendas have all but eliminated funding on the topic. However, in the form of unstructured natural language published daily by newspapers across the country, data abounds. We argue that this is the exact type of information that NLP is designed to organize, and the positive social impact of doing so would be substantial. We introduce the Gun Violence Database (GVDB), a new dataset of gun violence articles paired with NLP annotations. Our hope is that the GVDB will facilitate the adaptation of core NLP technologies to the domain of gun violence. In turn, we believe these NLP technologies can help overcome the data vacuum that is currently preventing productive discussion about gun violence and its possible solutions.

What we have:	Daily reports of gun violence, published as free text by local newspapers and TV stations.
What we need:	Structured, queryable database with one record per incident.

Information Retrieval: Find articles about gun violence.
Event Detection: Identify precise incident being reported.
Temporal Annotation: Pinpoint precise time of the event.
NER: Extract key locations and participants from the event.
Semantic Role Labeling: Relate participants to their role in the incident (e.g. shooter, victim).
With-document Coref: Resolve mentions to consistently model each participant throughout the event.
Semantic Parsing: Extract precise, detailed information about participants, e.g. race, age, and gender.
Cross-document Coref: Recognize mentions of the same shooter or victim appearing in different articles.
Event Coref: Identify articles reporting the same event, and resolve to a single database entry.

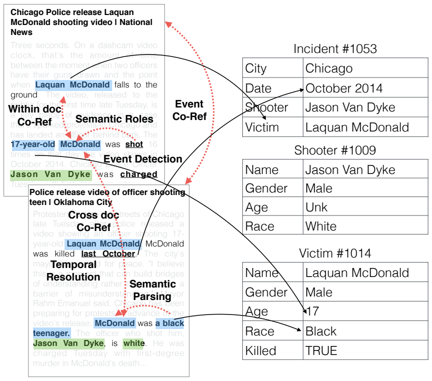

Chicago Police release Laquan McDonald shooting video I National News
Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed an officer behind bars. The 17-year-old McDonald was shot 16 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder
Within doc Co-Ref
Semantic Roles
Event Detection
Event Co-Ref
Police release video of officer shooting teen I Oklahoma City
Protesters took to the streets of Chicago late Tuesday after police released a video showing an officer shooting 17-year-old Laquan McDonald. McDonald was killed last October. The city's mayor called for peace. "I believe this is a moment that can build bridges of understanding rather than a barrier of misunderstanding," Mayor Rahm Emanuel said. Chicago has been preparing for protests in advance of the video's release. McDonald was a black teenager. The officer who shot him, Jason Van Dyke, is white. He was charged Tuesday with first-degree murder in McDonald's death…
Cross doc Co-Ref
Temporal Resolution
Semantic Parsing
Incident #1053
City | Chicago
Date | October 2014
Shooter | Jason Van Dyke
Victim | Laquan McDonald
Shooter #1009
Name | Jason Van Dyke
Gender | Male
Age | Unk
Race | White
Victim #1014
Name | Laquan McDonald
Gender | Male
Age | 17
Race | Black
Killed | TRUE

Figure 1: Turning daily news reports into usable data for public health and social science researchers is a textbook application of NLP technologies, and one that can have meaningful social impact.

2 Gun Violence's Data Problem

It is not difficult to motivate why gun violence is an important problem for research. Gun violence causes approximately 34,000 deaths in the US every year and more than twice as many injuries (FICAP, 2006), with violence especially high among young people and racial minorities (CDC, 2013).

The magnitude of the gun violence problem, the inherent gravity of the topic, and that fact that it inevitably leads to discussion of race, personal safety, and constitutional rights, makes the topic highly emotional and politically charged. Research into such hot-blooded topics stands to benefit immensely from data. In the past decade, machine learning researchers have championed data-driven decision making in place of oft-fallible human intuition. This approach has revolutionized the way we design and evaluate the effectiveness of business practices (Brynjolfsson et al., 2011; Kohavi et al., 2009), advertisements (Breese et al., 1998), and political campaigns (Issenberg, 2013). Gun violence policy should be no different. The problem is that researchers lack the data they need to answer the questions they want to ask. There is no single database[1] of gun violence incidents in the US, and the data that is available is mostly aggregated at the state level. Without locally-aggregated data, it is impossible to conduct meaningful studies of how firearm injury varies by community, a key step toward designing good policies for prevention (FICAP, 2006). However, for the past 25 years, research in this area has been, in the best case, massively underfunded (Roth et al., 1993) and in the worst case, actively blocked by federal legislation (Kassirer, 1995; Frankel, 2015; Bertrand, 2015). As a result, federal resources for gun violence research are orders of magnitude lower than is warranted (Branas et al., 2005), and there is no near-term likelihood of a federally-funded effort to collect detailed datasets to facilitate gun violence research.

Why NLP? Local newspapers and television stations report daily on gun injuries and fatalities. Many of these stories never make national news, but they represent precisely the kind of high-resolution data that epidemiologists need. The details of these reports could transform gun violence research if they were in a structured database, rather than spread across the text of thousands of web pages.

Replacing expensive, manual data entry with automated processing is exactly the type of problem that NLP is made to solve. In fact, the recent application of NLP tools to social science problems has generated a flurry of exciting and encouraging results. NLP has made novel contributions to the way scientists measure everything from income (Preoctiuc-Pietro et al., 2015b) to mental health (Preoctiuc-Pietro et al., 2015a; Schwartz et

[1] There are 13 national data systems in the U.S., managed by separate federal agencies. The National Violent Death Registry System, arguably the most organized effort, receives data from only 16 states. Most large-scale epidemiological studies sample information from only 100 Emergency Departments.

al., 2016; Choudhury et al., 2016), disease (Santillana et al., 2015; Ireland et al., 2015; Eichstaedt et al., 2015), and the quality of patient care (Nakhasi et al., 2016; Ranard et al., 2016).

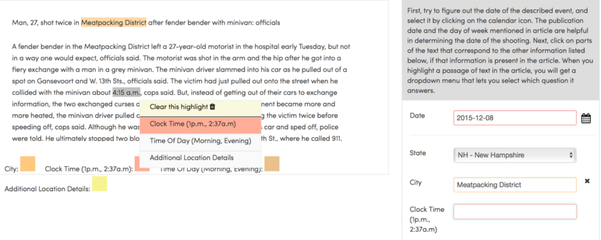
Text mining has promise for the study of gun violence, too (Bushman et al., 2016). However, most questions about gun violence are not easily answered using shallow analyses like topic models or word clusters. Epidemiologists want to know, for example, does gun ownership lead to increases in gun violence? Or, is there evidence of contagion in suicides, and if so, does the style of reporting on suicides affect the likelihood that others will commit suicide after the initial event? Answering these questions requires extracting precise information from text: identifying entities, their actions, and their attributes specifically and reliably.

We believe this level of depth is well within the reach of current NLP technology. The state-of-the-art tools that NLP researchers have been building and fine-tuning for decades are an ideal fit for the problem described. Nearly every step of this process, from retrieving articles about gun violence to correctly determining whether the phrase *14 year old girl* describes the victim or the shooter, has been studied as a core NLP problem in its own right (Figure 1). These NLP tools have the potential to make a marked difference for gun violence researchers.

3 The Gun Violence Database

In order to facilitate the adaptation of NLP tools for use in gun violence research, we introduce the Gun Violence Database[2] (GVDB), a dataset for training and evaluating the performance of NLP systems in the domain of gun violence. The GVDB is the result of a large crowdsourced annotation effort. This annotation is ongoing, and the GVDB will be regularly updated with new data and new layers of annotation, making it an interesting and challenging data set on which to evaluate state-of-the-art NLP tools.

Crowdsourced Annotation The GVDB is built and updated through a continuously running crowdsourced annotation pipeline. The pipeline consists of daily crawls of local newspapers and television websites from across the US. The crawled articles are automatically classified using a high-recall text classifier, and then manually vetted by humans to filter out false positives. So far, the GVDB contains 60K articles (∼49M words) describing incidents of gun violence, and is (sadly) growing at a rate of nearly 1,000 per day.



Man, 27, shot twice in Meatpacking District after fender bender with minivan: officials

A fender bender in the Meatpacking District left a 27-year-old motorist in the hospital early Tuesday, but not in a way one would expect, officials said. The motorist was shot in the arm and the hip after he got into a fiery exchange with a man in a grey minivan. The minivan driver slammed into his car as he pulled out of a spot on Gansevoort and W. 13th Sts., officials said. The victim had just pulled out onto the street when he collided with the minivan about 4:15 a.m., cops said. But, instead of getting out of their cars to exchange information, the two exchanged curses c nent became more and more heated, the minivan driver pulled c g the victim twice before speeding off, cops said. Although he was car and sped off, police were told. He ultimately stopped two blo th St., where he called 911.

Clear this highlight
Clock Time (1p.m., 2:37a.m)
Time Of Day (Morning, Evening)
Additional Location Details
City:
Clock Time (1p.m., 2:37a.m):
Time Of Day (Morning, Evening):
Additional Location Details:

First, try to figure out the date of the described event, and select it by clicking on the calendar icon. The publication date and the day of week mentioned in article are helpful in determining the date of the shooting. Next, click on parts of the text that correspond to the other information listed below, if that information is present in the article. When you highlight a passage of text in the article, you will get a dropdown menu that lets you select which question it answers.

Date: 2015-12-08

State: NH - New Hampshire

City: Meatpacking District

Clock Time (1p.m., 2:37a.m):

Figure 2: Annotation interface associates structured information (e.g. the time of day when the shooting occurred) with a specific span of text in the article.

Crowdsourced annotators then mark up the text of the articles with the key information we expect automated NLP systems to extract. In addition to classifying articles according to multiple binary dimensions (e.g. whether or not the shooting was intentional), annotators mark specific spans of the text which populate the database schema. For example, workers highlight the shooters, the victims, and the location.[3] These precise spans are stored in the database so that automated systems can be trained to reproduce the extracted information. Our annotation interface is shown in Figure 2.

At the time of writing, the GVDB contains 7,366 fully annotated articles (Table 1) coming from 1,512 US cities, and the database is continuing to grow. The latest version of the database will be maintained and available for download at `http://...`.

60,443	Articles reporting incidents of gun violence
7,366	Articles fully-annotated for IE
6,804	*w/ location information*
5,394	*w/ shooter/victim information*
4,143	*w/ temporal information*
1,666	*w/ weapon information*

Table 1: Current contents of the GVDB. Size and level of annotation is continually growing. See Forthcoming Extensions.

[2] `http://gun-violence.org/`

[3] See supplementary material for all extracted information and screenshots.

Current Baselines To establish a baseline level of performance, we run an off-the-shelf information extraction system on the 7,366 articles and measure precision and recall for identifying key information about the incidents. We use the Li et al. (2013) systems, which identifies a range of entities and events. We focus on the those events identified by the system which are relevant to the main fields in the GVDB schema.[4] We map the arguments of these events onto the corresponding database fields, e.g. the *agent* of the event corresponds to the GVDB's *shooter name*. Since the system identifies multiple such events per article, we count it as correct as long as one argument correctly matches the corresponding value in the GVDB (e.g. the system is correct as long as one extracted event has an *agent* which matches the GVDB's *shooter name* for that article). In addition, we run the Stanford CoreNLP TimeEx system (Chang and Manning, 2012) over the articles in order to identify the time of the reported incident.

We report the system's performance using both exact match against the gold annotation ("strict") as well as an approximate match, in which the system is correct if it is either a substring or a superstring of the gold annotation. E.g. if the victim name is *Sean Bolton*, the approximate metric will count both *Bolton* and *Officer Sean Bolton* as correct.

	Strict		Approx.	
	Prec.	Rec.	Prec.	Rec.
Date/Time	69.3%	66.9%	70.5%	68.1%
Location	19.9%	8.8%	30.8%	13.6%
Victim	10.2%	8.5%	59.5%	49.6%
Shooter	5.8%	3.9%	30.2%	20.1%
Weapon	2.1%	0.7%	36.8%	11.8%

Table 2: Performance of an off-the-shelf IE system on identifying key information about gun violence incidents from news articles. For "strict" vs. "approximate", see text.

While performance is high for certain structured types of information, like dates and times, fields like victim and shooter name are much less reliably identified. Furthermore, many key pieces of information in the GVDB, such as age and race, are not supported by the off-the-shelf system. These baselines are evidence that NLP systems have potential, but require some effort to make their output usable for downstream research. Our hope is that the GVDB will serve as the impetus for undertaking this effort.

Forthcoming Extensions The building of the GVDB is an ongoing effort, with new articles and deeper annotation being continuously added. We are currently adding approximately 300 new fully-annotated articles per day, while simultaneously enriching the annotation pipeline. The GVDB is soon to include annotation for event coreference, which will link articles describing the same incident, and cross-document coreference, which will link mentions of the same shooter/victim appearing in separate documents. In the future, the database will also include full within-document coreference annotation, with all mentions of a shooter/victim being flagged as such, and will incorporate visual data, so that within-article images are tagged with relevant information which may not be communicated by the text alone (e.g. race/approximate age).

4 Related Efforts

Several projects collect data about gun violence via newspaper teams (Boyle, 2013; Swaine et al., 2015) or volunteer crowds (Burghart, 2014; Wagner, 2014; Kirk and Kois, 2013). Perhaps the largest such effort is the Gun Violence Archive[5]. However, none are aimed at the eventual automation of the process. We believe that automating this data collection is key to keeping it scalable, consistent, and unbiased. Our focus is therefore on collecting data that is well-suited for training and evaluating NLP systems.

5 Conclusion

We believe that NLP researchers have the potential to significantly advance gun violence research. The shortage of data and funding for studying gun violence in America has severely limited the ability of scientists to have productive conversations about practical solutions. Applying core NLP technologies to local news reports of gun violence could transform raw text into structured, queryable data that public health researchers can use. We have introduced the Gun Violence Database, a new dataset of gun violence articles with rich NLP annotations which will support efforts on this new NLP task.

[4] Specifically, we focus on *Attack*, *Injure*, and *Die* events

[5] http://www.gunviolencearchive.org

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, October. Association for Computational Linguistics.

Natasha Bertrand. 2015. Congress quietly renewed a ban on gun-violence research. *Business Insider (July 7)*.

Andy Boyle. 2013. Mapping Chicago's shooting victims (Chicago Tribune), July.

Charles C Branas, Douglas J Wiebe, CW Schwab, and TS Richmond. 2005. Getting past the f word in federally funded public health research. *Injury prevention*, 11(3):191–191.

John S Breese, David Heckerman, and Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52.

Erik Brynjolfsson, Lorin M Hitt, and Heekyung Hellen Kim. 2011. Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

Brian D. Burghart. 2014. What I've learned from two years collecting data on police killings, August.

Brad J. Bushman, Katherine Newman, Sandra L. Calvert, Geraldine Downey, Mark Dredze, Michael Gottfredson, Nina G. Jablonski, Ann S. Masten, Calvin Morrill, Daniel B. Neill, Daniel Romer, and Daniel W. Webster. 2016. Youth violence: What we know and what we need to know. *American Psychologist*, 71(1):17–39, Jan.

CDC. 2013. Deaths: Final data for 2013. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 64(2).

Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *LREC*, pages 3735–3740.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Conference on Human Factors in Computing Systems (CHI)*.

Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

FICAP. 2006. *Firearm injury in the US.* Online Resource Book from The Firearm and Injury Center at Penn.

Todd C Frankel. 2015. Why the CDC still isn't researching gun violence, despite the ban being lifted two years ago. *The Washington Post (January 14)*.

Molly E Ireland, Qijia Chen, H Andrew Schwartz, Lyle H Ungar, and Dolores Albarracin. 2015. Action tweets linked to reduced county-level HIV prevalence in the United States: Online messages and structural determinants. *AIDS and Behavior*, pages 1–9.

Sasha Issenberg. 2013. How president Obama's campaign used big data to rally individual voters. *Technology Review*, 116(1):38–49.

Anne Kao and Steve R Poteet. 2007. *Natural language processing and text mining*. Springer Science & Business Media.

Jerome P Kassirer. 1995. A partisan assault on science–the threat to the CDC. *New England journal of medicine*, 333(12):793–794.

Chris Kirk and Dan Kois. 2013. How many people have been killed by guns since Newtown?, December.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria, August. Association for Computational Linguistics.

Atul Nakhasi, Sarah G Bell, Ralph J Passarella, Michael J Paul, Mark Dredze, and Peter J Pronovost. 2016. The potential of Twitter as a data source for patient safety. *Journal of Patient Safety*, Jan.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, Los Angeles, California, June. Association for Computational Linguistics.

Daniel Preoctiuc-Pietro, Maarten Sap, H Andrew Schwartz, and Lyle H Ungar. 2015a. Mental illness detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the*

Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, NAACL.

Daniel Preoctiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE*, 10(9), 09.

Benjamin L Ranard, Rachel M Werner, Tadas Antanavicius, H Andrew Schwartz, Robert J Smith, Zachary F Meisel, David A Asch, Lyle H Ungar, and Raina M Merchant. 2016. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs*, 35(4):697–705.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia, June. Association for Computational Linguistics.

Jeffrey A Roth, Albert J Reiss Jr, et al. 1993. *Understanding and preventing violence*, volume 1. National Academies Press.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 106–111, Boulder, Colorado, June. Association for Computational Linguistics.

Mauricio Santillana, Andre T. Nguyen, Mark Dredze, Michael J. Paul, Elaine Nsoesie, and John S. Brownstein. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLOS Computational Biology*.

H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, Michal Kosinski, Martin EP Seligman, and Lyle H. Ungar. 2016. Predicting Individual Well-Being Through the Language of Social Media. *Pacific Symposium on Biocomputing*, 21:516–527.

Amit Singhal. 2012. Introducing the knowledge graph: things, not strings. *Official Google Blog, May*.

Jon Swaine, Oliver Laughland, Jamiles Lartey, and Ciara McCarthy. 2015. The counted: People killed by police in the US (The Guardian), June.

Kyle Wagner. 2014. We're compiling every police-involved shooting in America. Help us., August.

William Yang Wang, Yashar Medhad, Dragomir Radev, and Amanda Stent. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA. ACL.

Derry Tanti Wijaya, Ndapandula Nakashole, and Tom Mitchell. 2015. A spousal relation begins with a deletion of engage and ends with an addition of divorce: Learning state changing verbs from Wikipedia revision history. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 518–523, Lisbon, Portugal, September. Association for Computational Linguistics.