Extracting structured information via automatic + human computation

Chris Callison-Burch
ccb@upenn.edu +1 267 909 2668
University of Pennsylvania
Department of Computer and Information Science
3330 Walnut Street, Philadelphia, PA 19104
Google Sponsors: Jakob Uszkoreit, Keith Hall
Google Contacts: Bill MacCartney, Brian Roark, Ashish Venugopal

Abstract

I propose to develop a methodology for extracting structured information from texts on the web using a combination of information retrieval, natural language processing, machine learning and crowdsourcing. I will demonstrate this methodology with a novel knowledge-base population task. I will develop a structured database of all incidents of gun violence in the United States that are reported in local newspapers and on local television stations' web sites. The methodology will be useful for researchers in relation extraction and semantic parsing, and the resulting database will be a valuable resource for epidemiologists and policy makers. The project will be built as part of my undergraduate crowdsourcing course (`crowdsourcing-class.org`), where it will be actively developed by 50 undergraduate students.

1 Motivation

Knowledge is represented in many ways. Sometimes it comes in structured formats like tables or relational databases bases. Often times it is conveyed in a less structured fashion through natural language. Structure makes it easier for us to draw inferences about data, to answer questions, or to perform better web searching. Google has improved search by building a knowledge graph that takes advantages of structured data like Freebase. I propose to develop a methodology for extracting structured knowledge from the web through a combination of automatic techniques (including machine learning, natural language processing and information retrieval/extraction) with human computation/crowdsourcing. The addition of human computation will allow us to reach a higher quality of data than is currently possible through fully-automatic techniques. To demonstrate this methodology, I will focus on a concrete task. I will build a structured database that details all reported incidents of gun violence in the United States.

This challenge task will serve as an illustration of the intellectual merit of the methodology as well as demonstrate the potential broader impact of being able to extract structured data from the web:

- Gun violence causes ≈33,000 deaths in the US every year and many more nonfatal injuries. Firearm injury is the fifth leading cause of years of potential life lost (YPLL), and it is second after motor vehicle accidents in terms of injury-related deaths (FICAP, 2006).
- There is no single database that details all or most US gun violence incidents. This stymies public health research, and prevents data-driven reasoning from being applied to policy creation.
- Local newspapers and television stations report on gun fatalities. The details of these reports would be valuable to epidemiologists if they were in a structured database, rather than spread across the text of thousands of web pages.

Although I focus on gun violence, the methodology that I develop in this project will be general. It could be applied to other events for which detailed global data is difficult to access, but which is described in a dispersed fashion on the web.

2 Proposed Work

I propose to establish a methodology that extracts structured data from the web using human computation as a primary component, augmenting ML, NLP, IR, and IE algorithms. Crowdsourcing enables the low-cost creation of high-quality data that can be used to train algorithms or to ensure that their output is correct. Here is a sketch of how to extract structured information about all gun violence incidents in the US:

1. Perform a daily web crawl of all local newspapers and TV stations. My students and I have collected 5 million articles from over 2,500 local newspapers.[1] These newspapers cover 50 states and 2,000 cities (Irvine et al., 2014).

2. Train a text classifier to predict whether articles describe gun violence or not. We collected a training set of 10,000 gun violence articles from a New York Times blog called the Gun Report.[2] We create a set of non-gun-violence articles by randomly sampling articles from our local newspaper corpus.

3. Classify all articles in the full set, and hire crowd workers to validate the classifiers output for articles above a certain threshold.

4. Run NLP tools like named entity recognizers over the validated texts to highlight potentially relevant elements of the articles for the structured database.

5. Hire crowd-workers to read the articles and answer questions to populate the database fields. For the gun violence database, fields include things like date and location of the incident, type of weapon, information about the shooter and victim, and other details about the incident.

6. Retrain the ML and NLP components on the manually labeled data. Train an IE system to pre-populate the database fields, initially to be edited by workers.

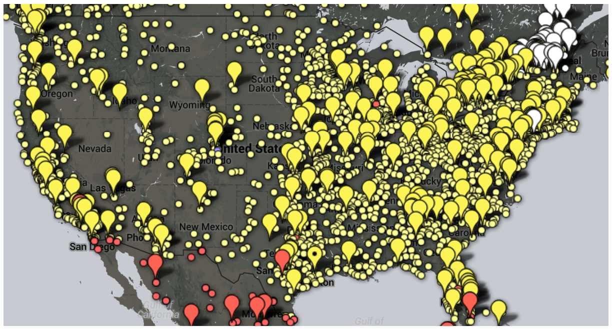

Figure 1: Our structured database will be extracted from local newspapers from 50 states and 2,000 cities gathered from NewspaperMap.

7. Iterate over steps 1-6, and measure the quality of the automatic predictions. The role of the humans in the annotation process can be reduced or eliminated when the quality of the automatic predictions reaches a sufficient level. Alternately, humans can vet everything, and the algorithms can be used to speed annotation.

The research questions include: how can we minimize the cost of extracting facts from text documents while ensuring accurate results? How can human annotation and automatic predictions best inform each other? Can we scale human-in-the-loop fact extraction from tens of thousands of events to hundreds of thousands or millions of events? Can relation extraction techniques for learning mappings from text to database fields be adapted from the Freebase/Wikipedia domains to other problems of interests to scientists and policy makers?

Our hybrid approach to information extraction will allow us to build a dataset that cannot currently be created using automatic methods alone. This fits with Google's mission of organizing the world's information. It will allow researchers in public health and epidemiology access to web-scale data in a query-able form that they can actually use. In addition, the resource will help advance technologies for building these kinds of resources fully automatically.

The expected outcomes of this project are twofold. First, we will build on existing research into relation extraction and semantic parsing (Mintz et al., 2009; Cai and Yates, 2013; Yao

[1] http://newspapermap.com

[2] http://nocera.blogs.nytimes.com/category/gun-report/

Condition	Total cases	NIH research awards
Cholera	373	101
Diphtheria	1,337	54
Polio	266	106
Rabies	55	59
Total for four diseases	**2,031**	**320**
Firearm injuries	>3,000,000	3

Table 1: Major NIH research awards and cumulative morbidity for select conditions in the US, 1973-2002. Reproduced from Branas et al. (2005).

and Van Durme, 2014) and into the cost-quality-time tradeoffs inherent to combining crowdsourcing and machine learning (Quinn and Bederson, 2011; Lin et al., 2014). Second, our proof-of-concept will result in a gun violence database that will be a useful artifact for public heath research.

3 Social Impact

The potential for social impact of the gun violence database is high. It could enable data-driven reasoning to be applied to a topic that is dominated by emotion. Research in this area is massively underfunded, and is actively blocked by federal legislation (Roth et al., 1993). Congress has prevented the Centers for Disease Control from funding research which may be used to "affect the passage of specific Federal, State, or local legislation intended to restrict or control the purchase or use of firearms" (Kassirer, 1995). Federal funding is not available to research this important topic. This extends to the National Institutes of Health (see Table 1). A privately funded, crowdsourced solution could have a huge impact.

The gun violence database would allow these important research questions to be answered without relying on government grants. The project would allow epidemiologists to take advantage of the enormous volume of natural language data available on the web, which they currently cannot process. At the same time, it would encourage NLP researchers to develop their technologies with respect to worthwhile applications with greater impact for society overall.

4 Data Policy

All of the data will be made freely available. The data collection efforts will be vetted by Penn's Institutional Review Board.

We will create the database schema in consultation with Douglas Weibe, a Professor of Epidemiology at Penn's School of Medicine who specializes in studying gun violence from a public health perspective. To date, epidemiologists have had difficulty systematically collecting data on gun violence. There is no centralized collection effort by the government, and the databases that do exist are incomplete and not updated in a timely fashion.[3]

5 Budget

We request $72,409. This will be used to fund 1 PhD student, plus $10,000 toward crowdsourcing costs. The student costs are based on University of Pennsylvania's standard rates for PhD students ($30,566 for tuition, $29,304 for student salary, $950 for a student computing fee), plus $1,500 for travel.

6 Results from Past Google Projects

I have received three Google faculty research awards in the past. In 2009, I was a co-PI with Miles Osborne on "The Babel Challenge: Translating the World's Languages." In 2011, I was a co-PI with Philip Resnik and Ben Bederson on "Translate the World: A Unified Framework for Crowdsourcing Translation," In 2013, I was the sole PI for "ParaGraph: Learning Paraphrases from Large, Diverse Data Sets." The first two awards focused on the low-cost creation of data for statistical machine translation systems. These awards dramatically influenced the direction of my research. Since receiving them, I have focused on crowdsourcing in my research, and I developed a course on Crowdsourcing and Human Computation (`crowdsourcing-class.org`).

I released several public data sets that I developed under these awards. This includes the paraphrase database (`paraphrase.org`), and several

[3] There are 13 national data systems in the U.S., managed by separate federal agencies. 16 states now report through the National Violent Death Registry System. Large-scale epidemiological studies sample information from 100 U.S. hospital Emergency Departments. There is no consistent standard for information like circumstances of firearm deaths.

crowdsourced translations set (translations of 10,000 individual words in each of 100 languages, and bilingual parallel corpora for six verb-final Indian languages with 0.5-1.5 million words in each language). Google's funding has also allowed me to develop an easy to use web site for performing translations on Amazon Mechanical Turk (`crowdtrans.com`).

References

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.

Qingqing Cai and Alexander Yates. 2013. Semantic parsing freebase: Towards open-domain semantic parsing. *Atlanta, Georgia, USA*, 30:328.

FICAP. 2006. *Firearm injury in the US*. Online Resource Book from The Firearm and Injury Center at Penn. http://www.uphs.upenn.edu/ficap/resourcebook/pdf/monograph.pdf.

Ann Irvine, Joshua Langfus, and Chris Callison-Burch. 2014. The American local news corpus. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May. European Language Resources Association.

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

Christopher H. Lin, Mausam, and Daniel S. Weld. 2014. To re(label), or not to re(label). In *HCOMP-2014*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Alexander J. Quinn and Benjamin B. Bederson. 2011. Human-machine hybrid computation. In *In CHI'11 Workshop*.

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

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with Freebase. In *Proceedings of ACL*.